

Non-Asymptotic and Second-Order Achievability Bounds for Coding With Side-Information

Shun Watanabe[†], Shigeaki Kuzuoka[‡], Vincent Y. F. Tan^{*}

Abstract

We present novel non-asymptotic or finite blocklength achievability bounds for three side-information problems in network information theory. These include (i) the Wyner-Ahlsvede-Körner (WAK) problem of almost-lossless source coding with rate-limited side-information, (ii) the Wyner-Ziv (WZ) problem of lossy source coding with side-information at the decoder and (iii) the Gel'fand-Pinsker (GP) problem of channel coding with noncausal state information available at the encoder. The bounds are proved using ideas from channel simulation and channel resolvability. Our bounds for all three problems improve on all previous non-asymptotic bounds on the error probability of the WAK, WZ and GP problems—in particular those derived by Verdú. Using our novel non-asymptotic bounds, we recover the general formulas for the optimal rates of these side-information problems. Finally, we also present achievable second-order coding rates by applying the multidimensional Berry-Esséen theorem to our new non-asymptotic bounds. Numerical results show that the second-order coding rates obtained using our non-asymptotic achievability bounds are superior to those obtained using existing finite blocklength bounds.

Index Terms

Source coding, channel coding, side-information, Wyner-Ahlsvede-Körner, Wyner-Ziv, Gel'fand-Pinsker, finite blocklength, non-asymptotic, second-order coding rates

I. INTRODUCTION

The study of *network information theory* [1] involves characterizing the optimal rate regions or capacity regions for problems involving compression and transmission from multiple sources to multiple destinations. Apart from a few special channels or source models, optimal rate regions and capacity regions for many network information theory problems are still not known. In this paper, we revisit three coding problems whose asymptotic rate characterizations are well known. These include

- The *Wyner-Ahlsvede-Körner* (WAK) problem of almost-lossless source coding with rate-limited (aka coded) side-information [2], [3],
- The *Wyner-Ziv* (WZ) problem of lossy source coding with side-information at the decoder [4], and
- The *Gel'fand-Pinsker* (GP) problem of channel coding with noncausal state information at the encoder [5].

These problems fall under the class of coding problems with *side-information*. That is, a subset of terminals has access to either a correlated source or the state of the channel. In most cases, this knowledge helps to strictly improve the rates of compression or transmission over the case where there is no side-information.

While the study of asymptotic characterizations of network information theory problems has been of key interest and importance for the past 50 years, it is important to analyze non-asymptotic (or finite blocklength) limits of various network information theory problems. This is because there may be hard constraints on decoding complexity or delay in modern, heavily-networked systems. The paper derives new non-asymptotic bounds on the error probability for the WAK and GP problems as well as the probability of excess distortion for the WZ problem. Our bounds improve on all existing finite blocklength bounds for these problems such as those in [6]. In addition, we use these bounds to recover known general formulas [7]–[10] and we also derive achievable second-order coding rates [11], [12] for these side-information problems.

[†]Department of Information Science and Intelligent Systems, University of Tokushima, Email: shun-wata@is.tokushima-u.ac.jp

[‡]Department of Computer and Communication Sciences, Wakayama University, Email: kuzuoka@ieee.org

^{*}Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR), and Department of Electrical and Computer Engineering (ECE), National University of Singapore (NUS), Emails: tanyfv@i2r.a-star.edu.sg, vtan@nus.edu.sg

Traditionally, the achievability proofs of the direct parts of each of these coding problems are common and involve a covering step, a packing step and finally the use of the Markov lemma [2] (also known as conditional typicality lemma in the book by El Gamal and Kim [1]). As such to prove tighter bounds, it is necessary to develop new proof techniques in place of the Markov lemma, covering and packing lemmas [1] and their non-asymptotic versions [6], [7]. These new techniques are based on the notion of *channel resolvability* [7], [13], [14] and *channel simulation* [15]–[19]. We use the former in the helper's code construction. Some historical remarks on the use of channel simulation to coding problems will be discussed in detail later in Section I-B.

To illustrate our idea at a high level, let us use the WAK problem as a canonical example of all three problems of interest. Recall that in the classical WAK problem, there is an independent and identically distributed (i.i.d.) joint source $P_{XY}^n(x^n, y^n) = \prod_{i=1}^n P_{XY}(x_i, y_i)$. The main source $X^n \sim P_X^n$ is to be reconstructed almost losslessly from rate-limited versions of both X^n and Y^n , where Y^n is a correlated random variable regarded as side-information or helper. See Fig. 1. The compression rates of X^n and Y^n are denoted as R_1 and R_2 respectively. The optimal rate region is the set of rate pairs (R_1, R_2) for which there exists a *reliable* code, that is one whose error probability can be made arbitrarily small with increasing blocklengths. WAK [2], [3] showed that the optimal rate region is

$$R_1 \geq H(X|U), \quad R_2 \geq I(U; Y) \quad (1)$$

for some $P_{U|Y}$. For the direct part, the helper encoder compresses the side-information and transmits a description represented by U^n . By the covering lemma [1], this results in the rate constraint $R_2 \geq I(U; Y)$. The main encoder then uses binning [20] as in the achievability proof of the Slepian-Wolf theorem [21] to help the decoder recover X given the description U . This results in the rate constraint $R_1 \geq H(X|U)$. The main idea in our proof of a new non-asymptotic upper bound on the error probability for the WAK problem is that, mixed over some common randomness of arbitrarily large cardinality, the joint distribution of (U, Y) (in the one-shot notation) is close in the variational distance sense to (\hat{U}, Y) , where \hat{U} designates the chosen auxiliary codeword (found classically via joint typicality encoding). As a result, by monotonicity and the data-processing lemma for the variational distance, it can be shown that the joint distribution of (U, Y, X) is also close to (\hat{U}, Y, X) . This means that in the asymptotic (n -fold i.i.d. repetition) setting, the triple (\hat{U}, Y, X) is jointly typical with high probability. This technique thus circumvents the need to use the so-called piggyback coding lemma (PBL) and the Markov lemma [2] which result in much poorer estimates on the error probability.

A. Main Contributions

We now describe the three main contributions in this paper.

Our first main contribution in this paper is to show improved bounds on the probabilities of error for WAK, WZ and GP coding. We briefly describe the form of the bound for WAK coding here. The primary part of the new upper bound on the error probability $P_e(\Phi)$ for WAK coding depends on two positive constants γ_b and γ_c and is essentially given by

$$P_e(\Phi) \lesssim \Pr(\mathcal{E}_c \cup \mathcal{E}_b) \quad (2)$$

where the *covering error* is

$$\mathcal{E}_c := \left\{ \log \frac{P_{Y|U}(U|Y)}{P_Y(Y)} \geq \gamma_c \right\} \quad (3)$$

and the *binning error* is

$$\mathcal{E}_b := \left\{ \log \frac{1}{P_{X|U}(X|U)} \geq \gamma_b \right\}. \quad (4)$$

The notation \lesssim is not meant to be precise and, in fact, we are dropping several residual terms that do not contribute to the second-order coding rates in the n -fold i.i.d. setting if γ_b and γ_c are chosen appropriately. This result is stated precisely in Theorem 14. From (2), we deduce that in the n -fold i.i.d. setting, if we choose γ_c and γ_b to be fixed numbers that are strictly larger than the mutual information $I(U; Y)$ and the conditional entropy $H(X|U)$ respectively, we are guaranteed that the error probability $P_e(\Phi)$ decays to zero. This follows from Khintchine's law of large numbers [7, Ch. 1]. Thus, we recover the direct part of WAK's result. In fact, we can take this one step further (Theorem 21) to obtain an achievable *general formula* (in the sense of Verdú-Han [7], [22]) for the WAK problem with general source [7, Ch. 1]. This was previously done by Miyake-Kanaya [8] but their derivation

is based on a different non-asymptotic formula more akin to Wyner's PBL. Also, since we have the freedom to design γ_c and γ_b as sequences instead of fixed positive numbers, if we let them be $O(\frac{1}{\sqrt{n}})$ -larger than $I(U; Y)$ and $H(X|U)$, then the error probability is smaller than a prescribed constant depending on the implied constants in the $O(\cdot)$ -notations. This follows from the multivariate Berry-Esséen theorem [23]. This bound is useful because it is a union of two events and \mathcal{E}_c and \mathcal{E}_b are both information spectrum [7] events which are easy to analyze.

Secondly, the preceding discussion shows that the bound in (2) also yields an achievable second-order coding rate [11], [12]. However, unlike in the point-to-point setting [11], [12], [24], the achievable second-order coding rate is expressed in terms of a so-called *dispersion matrix* [25]. We can easily show that if $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ is the set of all rate pairs (R_1, R_2) for which there exists a length- n WAK code with error probability not exceeding $\varepsilon > 0$ (i.e., the (n, ε) -optimal rate region), then for any $P_{U|Y}$ and all n sufficiently large, the set

$$\begin{bmatrix} I(U; Y) \\ H(X|U) \end{bmatrix} + \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + O\left(\frac{\log n}{n}\right) \mathbf{1}_2 \quad (5)$$

is an inner bound to $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$. In (5), $\mathcal{S}(\mathbf{V}, \varepsilon) \subset \mathbb{R}^2$ denotes the analogue of the Q^{-1} function [25] and it depends on the covariance matrix of the so-called information-entropy density vector

$$\left[\log \frac{P_{Y|U}(U|Y)}{P_Y(Y)} \quad \log \frac{1}{P_{X|U}(X|U)} \right]^T. \quad (6)$$

The precise statement for the second-order coding rate for the WAK problem is given in Theorem 24. We see from (5) that for a fixed test channel $P_{U|Y}$, the redundancy at blocklength n in order to achieve an error probability $\varepsilon > 0$ is governed by the term $\frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}}$. The pre-factor of this term $\mathcal{S}(\mathbf{V}, \varepsilon)$, is likened to the *dispersion* [24], [26]–[28], and depends not only the variances of the information and entropy densities but also their correlations.

Thirdly, we note that the same flavour of non-asymptotic bounds and second-order coding rates hold verbatim for the WZ and GP problems. In addition, since the canonical rate-distortion problem [29] is a special case of the WZ problem, we show that our non-asymptotic achievability bound for the WZ problem, when suitably specialized, yields the correct dispersion for lossy source coding [27], [28]. We do so using two methods: (i) the method of types [30] and (ii) results involving the D -tilted information [28]. Finally, we not only improve on the existing bounds for the GP problem [6], [10], but we also consider an almost sure cost constraint on the channel input.

B. Related Work

Wyner [2] and Ahlswede-Körner [3] were the first to consider and solve the problem of almost-lossless source coding with coded side information. Weak converses were proved in [2], [3] and a strong converse was proved in [31] using the “blowing-up lemma”. An information spectrum characterization was provided by Miyake and Kanaya [8] and Kuzuoka [32] leveraged on the non-asymptotic bound which can be extracted from [8] to derive the redundancy for the WAK problem. Verdú [6] strengthened the non-asymptotic bound and showed that the error probability for the WAK problem is essentially bounded as

$$P_e(\Phi) \lesssim \Pr(\mathcal{E}_c) + \Pr(\mathcal{E}_b), \quad (7)$$

which is the result upon using the union bound on our bound in (2). Again, we used the notation \lesssim to mean that the residual terms do not affect the second-order coding rates. Bounds on the error exponent were derived by Kelly-Wagner [33].

Wyner and Ziv [4] derived the rate-distortion function for lossy source coding with decoder side-information. However, they do not consider the probability of excess distortion. Rather, the quantity of interest is the expected distortion and, more precisely, they considered the constraint that the asymptotic expected distortion is below a distortion threshold $D > 0$. The generalization of the WZ problem for general correlated sources was considered by Iwata and Muramatsu [9] who showed that the general WZ function can be written as a difference of a limit superior in probability and a limit inferior in probability, reflecting the covering and packing components in the classical achievability proof. Bounds on the error exponent were provided by Kelly-Wagner [33].

The problem of channel coding with noncausal random state information was solved by Gel'fand and Pinsker [5]. Subsequent work by Costa showed that, remarkably, there is no rate loss in the Gaussian case [34]. This is done by choosing the auxiliary random variable to be a linear combination of the channel input and the state. A general

formula for the GP problem (with general channel and general state) was provided by Tan [10]. An achievable error exponent was derived by Moulin and Wang [35]. Tyagi and Narayan [36] proved the strong converse for this problem and used it to derive a sphere-packing bound. For both the WZ and GP problems, Verdú [6] used generalizations of the (asymptotic) packing and covering lemmas in El Gamal and Kim [1] to derive non-asymptotic bounds on the probability of excess distortion (for WZ) and the average error probability (for GP). However, they yield slightly worse second-order rates because the main part of the bound is a sum of two or three probabilities as in (7), rather than the probability of the union as in (2).

In our work, we derive tight non-asymptotic bounds by using ideas from channel resolvability [13] [7, Ch. 6] and channel simulation [16], [17]¹ to replace the covering part and Markov lemma. For a given channel W and an input distribution, *channel resolvability* concerns the approximation of the output distribution with as small amount of randomness at the input as possible. It was shown by Han and Verdú [13] that this problem is closely connected to channel coding and channel identification. Hayashi also studied the channel resolvability problem [14] and derived a non-asymptotic formula that is different from Han and Verdú's. We will leverage on a key lemma in Hayashi [14] to derive our finite blocklength bounds.

In [16], [17], Bennett *et al.* proposed a problem to simulate a channel by the aid of common randomness. An application of the channel simulation to simulate the test channel in the rate-distortion problem was first investigated by Winter [18], and then extensively studied mainly in the field of the quantum information (eg. [15], [38], [39]). Cuff investigated the trade-off between the rates of the message and common randomness for the channel simulation [19]. In these literatures, the channel resolvability is implicitly or explicitly used as a building block of the channel simulation. Although the ideas to use the channel simulation instead of the Markov lemma is motivated by above mentioned literatures, we stress that the derivations of tight non-asymptotic bounds as in (2) are not straightforward applications of the channel simulation and are highly nontrivial, which are technical contributions of this paper. Indeed, our code construction of the channel simulation is slightly different from the literatures, and we also introduce some bounding techniques that have not appeared in any literatures.

In [40], Yassaee *et al.* proposed an alternative approach for channel simulation, in which they essentially used the (multi-terminal version of) intrinsic randomness [7, Ch. 2] instead of channel resolvability. Although their approach can be also used to replace the Markov lemma, it is not yet clear whether our bound can be also derived from the approach in [40]. One of difficulties to apply the approach in [40] for non-asymptotic analysis is that the amount of common randomness that can be used in the channel simulation is limited by the randomness of sources involved in a coding problem, which is not the case with the approach using the channel resolvability. More precisely, the channel simulation errors in both approach involve terms stemmed from the amounts of common randomness. In the channel resolvability approach, we can make the amount of common randomness arbitrarily large, and thus make that term arbitrarily small, which is not the case with the approach in [40]. See our Theorems 14, 17 and 19.

Our main motivation in this work is to derive tight finite blocklength bounds on the error probability (or probability of excess distortion). We are also interested in second-order coding rates. The study of the asymptotic expansion of the logarithm of the maximum number codewords that are achievable for n uses a channel with maximum error probability no larger than ε was first done by Strassen [41]. This was re-popularized in recent times by Kontoyiannis [42], Baron-Khojastepour-Baraniuk [43], Hayashi [11], [12], and Polyanskiy-Poor-Verdú [24] among others. Other notable works in this area include those by Nomura-Han [44] for resolvability, Kostina-Verdú [28] for lossy source coding and Wang-Ingber-Kochman [26] for joint source channel coding. Second-order analysis for network information theory problems were considered in Tan and Kosut [25] as well as other authors [45]–[48]. However, this is the first work that considers second-order rates for problems with side-information that are not straightforward extensions of other known results.

C. Paper Organization

In Section II, we state our notation and formally define the three coding problems with side-information. We then review existing asymptotic, non-asymptotic and error exponent-type results in Section III. In Section IV, we state our new non-asymptotic, channel-simulation-type bounds for the three problems. We then use these bounds to re-derive (the direct parts of the) known general formulas [8], [10] in Section V. Following that, we present achievable

¹Steinberg and Verdú also studied the channel simulation problem [37]. However, their problem formulation is slightly different from the one in [16], [17].

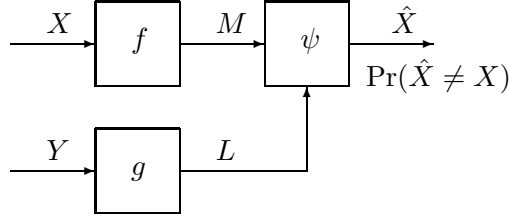


Fig. 1. Illustration of the WAK problem

second-order coding rates for these coding problems. We will see that just as in the Slepian-Wolf setting [25], [47], the dispersion is in fact a matrix. In Section VII, we show via numerical examples that our non-asymptotic bounds lead to larger (n, ε) -rate regions compared with [6]. Concluding remarks and directions for future work are provided Section VIII. To ensure that the main ideas are seamlessly communicated in the main text, we relegate all proofs to the appendices.

II. PRELIMINARIES

In this section, we introduce our notation and recall the WAK, WZ and GP problems.

A. Notations

Random variables (e.g., X) and their realizations (e.g., x) are in capital and lower case respectively. All random variables take values in some alphabets which are denoted in calligraphic font (e.g., \mathcal{X}). The cardinality of \mathcal{X} , if finite, is denoted as $|\mathcal{X}|$. Let the random vector $X^n := (X_1, \dots, X_n)$ and similarly for a realization $x^n = (x_1, \dots, x_n)$. The set of all distributions supported on alphabet \mathcal{X} is denoted as $\mathcal{P}(\mathcal{X})$. We will at times use the method of types [30]. The joint distribution induced by a marginal distribution $P \in \mathcal{P}(\mathcal{X})$ and a channel law $V : \mathcal{X} \rightarrow \mathcal{Y}$ is denoted interchangeably as $P \times V$ or PV . This should be clear from the context.

For a sequence $x^n = (x_1, \dots, x_n) \in \mathcal{X}^n$ in which $|\mathcal{X}|$ is finite, its *type* or *empirical distribution* is the probability mass function $P(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x = x_i\}$. The set of types with denominator n supported on alphabet \mathcal{X} is denoted as $\mathcal{P}_n(\mathcal{X})$. The *type class* of P is denoted as $\mathcal{T}_P := \{x^n \in \mathcal{X}^n : x^n \text{ has type } P\}$. For a sequence $x^n \in \mathcal{T}_P$, the set of sequences $y^n \in \mathcal{Y}^n$ such that (x^n, y^n) has joint type $PV = P(x)V(y|x)$ is the *V-shell* $\mathcal{T}_V(x^n)$. Let $\mathcal{V}_n(\mathcal{Y}; P)$ be the family of stochastic matrices $V : \mathcal{X} \rightarrow \mathcal{Y}$ for which the *V-shell* of a sequence of type $P \in \mathcal{P}_n(\mathcal{X})$ is not empty. Information-theoretic quantities are denoted in the usual way. For example, $I(X; Y)$ and $I(P, V)$ denote the mutual information where the latter expression makes clear that the joint distribution of (X, Y) is PV . All logarithms are with respect to base 2 so information quantities are measured in bits.

The multivariate normal distribution with mean μ and covariance matrix Σ is denoted as $\mathcal{N}(\mu, \Sigma)$. The complementary Gaussian cumulative distribution function $Q(t) := \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ and its inverse is denoted as $Q^{-1}(\varepsilon) := \min\{t \in \mathbb{R} : Q(t) \leq \varepsilon\}$. Finally, $|z|^+ := \max\{z, 0\}$ and $\mathbf{1}\{x \in \mathcal{A}\} = 1$ if $x \in \mathcal{A}$ and 0 otherwise.

B. The Wyner-Ahlsvede-Körner (WAK) Problem

In this section, we recall the WAK problem of lossless source coding with coded side-information [2], [3]. Let us consider a correlated source (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$ and having joint distribution P_{XY} . Throughout, X , a discrete random variable, is the main source while Y is the helper or side-information. The WAK problem involves reconstructing X losslessly given rate-limited (or coded) versions of both X and Y . See Fig. 1.

Definition 1. A (possibly stochastic) source coding with side-information code or Wyner-Ahlsvede-Körner (WAK) code $\Phi = (f, g, \psi)$ is a triple of mappings that includes two encoders $f : \mathcal{X} \rightarrow \mathcal{M}$ and $g : \mathcal{Y} \rightarrow \mathcal{L}$ and a decoder $\psi : \mathcal{M} \times \mathcal{L} \rightarrow \mathcal{X}$. The error probability of the WAK code Φ is defined as

$$P_e(\Phi) := \Pr\{X \neq \psi(f(X), g(Y))\}. \quad (8)$$

In the following, we may call f as the main encoder and g the helper.

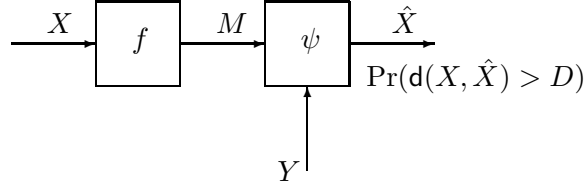


Fig. 2. Illustration of the WZ problem with probability of excess distortion criterion

In Section VI, we consider n -fold i.i.d. extensions of X and Y , denoted as X^n and Y^n . In this case, we use the subscript n to specify the blocklength, i.e., the code is $\Phi_n = (f_n, g_n, \psi_n)$ and the compression index sets are $\mathcal{M}_n = f_n(\mathcal{X}^n)$ and $\mathcal{L}_n = g_n(\mathcal{Y}^n)$. In this case, we can define the pair of rates of the code Φ_n as

$$R_1(\Phi_n) := \frac{1}{n} \log |\mathcal{M}_n|, \quad (9)$$

$$R_2(\Phi_n) := \frac{1}{n} \log |\mathcal{L}_n|. \quad (10)$$

Definition 2. The (n, ε) -optimal rate region for the WAK problem $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ is defined as the set of all pairs of rates (R_1, R_2) for which there exists a blocklength- n WAK code Φ_n with rates at most (R_1, R_2) and with error probability not exceeding ε . In other words,

$$\mathcal{R}_{\text{WAK}}(n, \varepsilon) := \left\{ (R_1, R_2) \in \mathbb{R}_+^2 : \exists \Phi_n \text{ s.t. } \frac{1}{n} \log |\mathcal{M}_n| \leq R_1, \frac{1}{n} \log |\mathcal{L}_n| \leq R_2, P_e(\Phi_n) \leq \varepsilon \right\} \quad (11)$$

We also define the asymptotic rate regions

$$\mathcal{R}_{\text{WAK}}(\varepsilon) := \text{cl} \left[\bigcup_{n \geq 1} \mathcal{R}_{\text{WAK}}(n, \varepsilon) \right], \quad (12)$$

$$\mathcal{R}_{\text{WAK}} := \bigcap_{0 < \varepsilon < 1} \mathcal{R}_{\text{WAK}}(\varepsilon). \quad (13)$$

where cl denotes set closure in \mathbb{R}^2 .

In the following, we will provide an inner bound to $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ that improves on inner bounds that can be derived from previously obtained non-asymptotic bounds on $P_e(\Phi_n)$ [6], [32].

C. The Wyner-Ziv (WZ) Problem

In this section, we recall the WZ problem of lossy source coding with full side-information at the decoder [4]. Here, as in the WAK problem, we have a correlated source (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$ and having joint distribution P_{XY} . Again, X is the main source and Y is the helper or side-information. Neither X nor Y has to be a discrete random variable. Unlike the WAK problem, it is not required to reconstruct X exactly, rather a distortion D between X and its reproduction \hat{X} is allowed. Let $\hat{\mathcal{X}}$ be the reproduction alphabet and let $d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow [0, \infty)$ be a bounded distortion measure such that for every $x \in \mathcal{X}$ there exists a $\hat{x} \in \hat{\mathcal{X}}$ such that $d(x, \hat{x}) = 0$ and $\max_{x, \hat{x}} d(x, \hat{x}) = D_{\max} < \infty$. See Fig. 2.

Definition 3. A (possibly stochastic) lossy source coding with side-information or Wyner-Ziv (WZ) code $\Phi = (f, \psi)$ is a pair of mappings that includes an encoder $f : \mathcal{X} \rightarrow \mathcal{M}$ and a decoder $\psi : \mathcal{M} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$. The probability of excess distortion for the WZ code Φ at distortion level D is defined as

$$P_e(\Phi; D) := \Pr\{d(X, \psi(f(X), Y)) > D\}. \quad (14)$$

We will again consider n -fold extensions of X and Y , denoted as X^n and Y^n in Section VI. The code is indexed by the blocklength as $\Phi_n = (f_n, \psi_n)$. Furthermore, the compression index set is denoted as $\mathcal{M}_n = f_n(\mathcal{X}^n)$. The rate of the code Φ_n is defined as

$$R(\Phi_n) := \frac{1}{n} \log |\mathcal{M}_n|. \quad (15)$$

The distortion between two length- n sequences $x^n \in \mathcal{X}^n$ and $\hat{x}^n \in \hat{\mathcal{X}}^n$ is defined as

$$d_n(x^n, \hat{x}^n) := \frac{1}{n} \sum_{i=1}^n d(x_i, \hat{x}_i). \quad (16)$$

Definition 4. The (n, ε) -Wyner-Ziv rate-distortion region $\mathcal{R}_{\text{WZ}}(n, \varepsilon) \subset \mathbb{R}_+^2$ is the set of all rate-distortion pairs (R, D) for which there exists a blocklength- n WZ code Φ_n at distortion level D with rate at most R and probability of excess distortion not exceeding ε . In other words,

$$\mathcal{R}_{\text{WZ}}(n, \varepsilon) := \left\{ (R, D) \in \mathbb{R}_+^2 : \exists \Phi_n \text{ s.t. } \frac{1}{n} \log |\mathcal{M}_n| \leq R, P_e(\Phi_n; D) \leq \varepsilon \right\} \quad (17)$$

We also define the asymptotic rate-distortion regions

$$\mathcal{R}_{\text{WZ}}(\varepsilon) := \text{cl} \left[\bigcup_{n \geq 1} \mathcal{R}_{\text{WZ}}(n, \varepsilon) \right], \quad (18)$$

$$\mathcal{R}_{\text{WZ}} := \bigcap_{0 < \varepsilon < 1} \mathcal{R}_{\text{WZ}}(\varepsilon). \quad (19)$$

The (n, ε) -Wyner-Ziv rate-distortion function $R_{\text{WZ}}(n, \varepsilon, D)$ is defined as

$$R_{\text{WZ}}(n, \varepsilon, D) := \inf \{ R : (R, D) \in \mathcal{R}_{\text{WZ}}(n, \varepsilon) \} \quad (20)$$

We also define the asymptotic rate-distortion functions

$$R_{\text{WZ}}(\varepsilon, D) = \inf \{ R : (R, D) \in \mathcal{R}_{\text{WZ}}(\varepsilon) \} \quad (21)$$

$$R_{\text{WZ}}(D) = \lim_{\varepsilon \rightarrow 0} R_{\text{WZ}}(\varepsilon, D) \quad (22)$$

Note that the use of the limit (as opposed to the limit superior or limit inferior) in (22) is justified because $R_{\text{WZ}}(\varepsilon, D)$ is, from its definition, monotonically non-increasing in ε . In the sequel, we will provide an inner bound to $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$ and thus an upper bound on $R_{\text{WZ}}(n, \varepsilon, D)$ by appealing to a new non-asymptotic upper bound on the probability of excess distortion $P_e(\Phi_n; D)$. In addition, note that if $Y = \emptyset$, i.e., side-information is not available, this reduces to the point-to-point rate-distortion (lossy source coding) problem.

Conventionally [1], [4], the WZ problem is stated not with the probability of excess distortion criterion but with the *average fidelity criterion*. That is, the requirement that $P_e(\Phi_n; D) \rightarrow 0$ (implicit in (22)) is replaced by

$$\limsup_{n \rightarrow \infty} \mathbb{E}[d_n(X^n, \psi_n(f_n(X^n), Y^n))] \leq D. \quad (23)$$

D. The Gel'fand-Pinsker (GP) Problem

In the previous two subsections, we dealt exclusively with source coding problems, either lossless (WAK) or lossy (WZ). In this section, we review the setup of the GP problem [5] which involves channel coding with noncausal state information at the encoder. It is the dual to the WZ problem [49]. In this problem, there is a state-dependent channel $W : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{Y}$ and a random variable representing the state S with distribution P_S taking values in some set \mathcal{S} . A message M chosen uniformly at random from \mathcal{M} is to be sent and the encoder has information about which message is to be sent as well as the channel state information S , which is known *noncausally*. (Noncausality only applies when the blocklength is larger than 1.) It is assumed that the message and the state are independent. Let $g : \mathcal{X} \rightarrow [0, \infty)$ be some cost function. The encoder f encodes the message and state into a codeword (channel input) $X = f(M, S)$ that satisfies the cost constraint

$$g(X) \leq \Gamma, \quad (24)$$

for some $\Gamma \geq 0$ with high probability. See precise definition/requirement in (25) as well as Proposition 1. The decoder receives the channel output $Y | \{X = x, S = s\} \sim W(\cdot | x, s)$ and decides which message was sent via a decoder $\psi : \mathcal{Y} \rightarrow \mathcal{M}$. See Fig. 3. More formally, we have the following definition.

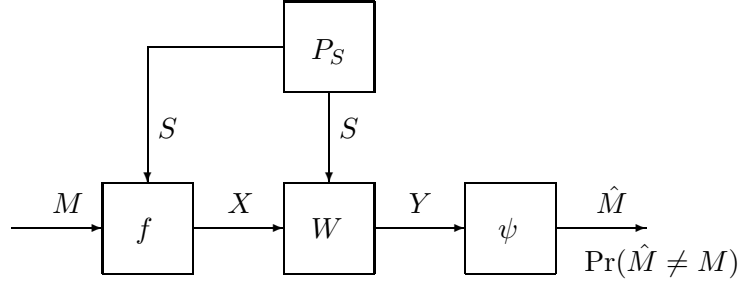


Fig. 3. Illustration of the GP problem

Definition 5. A (possibly stochastic) code for the channel coding problem with noncausal state information or Gel'fand-Pinsker (GP) code $\Phi = (f, \psi)$ is a pair of mappings that includes an encoder $f : \mathcal{M} \times \mathcal{S} \rightarrow \mathcal{X}$ and a decoder $\psi : \mathcal{Y} \rightarrow \mathcal{M}$. The average probability of error for the GP code is defined as

$$P_e(\Phi; \Gamma) := \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{S}} P_S(s) \sum_{y \in \mathcal{Y}} W(y|f(m, s), s) \mathbf{1} \{g(f(m, s)) > \Gamma \cup y \in \mathcal{Y} \setminus \psi^{-1}(m)\}. \quad (25)$$

More simply, $P_e(\Phi; \Gamma) = \Pr(\{g(f(M, S)) > \Gamma\} \cup \{\hat{M} \neq M\})$ where M is uniform on \mathcal{M} and independent of $S \sim P_S$, $\hat{M} := \psi(Y)$ and Y is the random variable whose conditional distribution given $M = m$ and $S = s$ is $W(\cdot | f(m, s), s)$.

The following proposition, which will be proved in Appendix A, guarantees that we can always convert a code in the sense of Definition 5 into a code in the sense of an almost sure cost constraint.

Proposition 1 (Expurgated Code). *Let the set of admissible inputs in \mathcal{X} be*

$$\mathcal{T}_g^{\text{GP}}(\Gamma) := \{x \in \mathcal{X} : g(x) \leq \Gamma\}. \quad (26)$$

For any (stochastic) encoder $P_{X|MS}$ (this plays the role of f in Definition 5) and decoder $P_{\hat{M}|Y}$ (this plays the role of ψ in Definition 5), there exists an encoder $\tilde{P}_{X|MS}$ such that

$$\tilde{P}_X(\mathcal{T}_g^{\text{GP}}(\Gamma)) = 1 \quad (27)$$

and

$$\tilde{P}_{MSXY\hat{M}}[m \neq \hat{m}] \leq P_{MSXY\hat{M}}[g(x) > \Gamma \cup m \neq \hat{m}], \quad (28)$$

where $P_{MSXY\hat{M}} := P_M P_S P_{X|MS} W P_{\hat{M}|Y}$ and $\tilde{P}_{MSXY\hat{M}} := P_M P_S \tilde{P}_{X|MS} W P_{\hat{M}|Y}$.

From Proposition 1, noting that $P_e((P_{X|MS}, P_{\hat{M}|Y}); \Gamma) = P_{MSXY\hat{M}}[g(x) > \Gamma \cup m \neq \hat{m}]$, we see that the constraint in (24) is equivalent to $g(X) \leq \Gamma$ almost surely (implied by (27)). For the purposes of deriving channel simulation-based bounds in Section IV-C, it is easier to work with the error criterion in (25) so we adopt Definition 5.

In order to obtain achievable second-order coding rates for the GP problem, we consider n -fold i.i.d. extensions of the channel and state. Hence, for every (s^n, x^n, y^n) , we have $W^n(y^n|x^n, s^n) = \prod_{i=1}^n W(y_i|x_i, s_i)$ and the state S^n evolves in a stationary, memoryless fashion according to P_S . For blocklength n , the code and message set are denoted as $\Phi_n = (f_n, \psi_n)$ and \mathcal{M}_n respectively. The cost function is denoted as $g_n : \mathcal{X}^n \rightarrow [0, \infty)$ and is defined as the average of the per-letter costs, i.e.,

$$g_n(x^n) := \frac{1}{n} \sum_{i=1}^n g(x_i) \quad (29)$$

For example, in the Gaussian GP problem (which is also known as *dirty paper coding* [34]), $g(x) = x^2$. This corresponds to a power constraint and Γ is the upper bound on the permissible power. The rate of the code is the normalized logarithm of the number of messages, i.e.,

$$R(\Phi_n) := \frac{1}{n} \log |\mathcal{M}_n|. \quad (30)$$

Definition 6. The (n, ε) -GP capacity-cost region $\mathcal{C}_{\text{GP}}(n, \varepsilon) \subset \mathbb{R}_+^2$ is the set of all rate-cost pairs (R, Γ) for which there exists a blocklength- n GP code Φ_n with cost not exceeding Γ , with rate at least R and probability of error not exceeding ε . In other words,

$$\mathcal{C}_{\text{GP}}(n, \varepsilon) := \left\{ (R, \Gamma) \in \mathbb{R}_+^2 : \exists \Phi_n \text{ s.t. } \frac{1}{n} \log |\mathcal{M}_n| \geq R, P_e(\Phi_n; \Gamma) \leq \varepsilon \right\}. \quad (31)$$

We also define the asymptotic capacity-cost regions

$$\mathcal{C}_{\text{GP}}(\varepsilon) := \text{cl} \left[\bigcup_{n \geq 1} \mathcal{C}_{\text{GP}}(n, \varepsilon) \right], \quad (32)$$

$$\mathcal{C}_{\text{GP}} := \bigcap_{0 < \varepsilon < 1} \mathcal{C}_{\text{GP}}(\varepsilon). \quad (33)$$

The (n, ε) -capacity-cost function $C_{\text{GP}}(n, \varepsilon, \Gamma)$ is defined as

$$C_{\text{GP}}(n, \varepsilon, \Gamma) := \sup \{ R : (R, \Gamma) \in \mathcal{C}_{\text{GP}}(n, \varepsilon) \} \quad (34)$$

We also define the asymptotic capacity-cost functions

$$C_{\text{GP}}(\varepsilon, \Gamma) := \sup \{ R : (R, \Gamma) \in \mathcal{C}_{\text{GP}}(\varepsilon) \} \quad (35)$$

$$C_{\text{GP}}(\Gamma) := \lim_{\varepsilon \rightarrow 0} C_{\text{GP}}(\varepsilon, \Gamma) \quad (36)$$

If the cost constraint (24) is absent (i.e., every codeword in \mathcal{X}^n is admissible), we will write $C_{\text{GP}}(n, \varepsilon)$ instead of $C_{\text{GP}}(n, \varepsilon, \infty)$, $P_e(\Phi_n)$ instead of $P_e(\Phi_n; \infty)$ and so on.

Once again, the limit in (36) exists because the function $C_{\text{GP}}(\varepsilon, \Gamma)$ is monotonically non-decreasing in ε . In the sequel, we will provide a lower bound on $C_{\text{GP}}(n, \varepsilon, \Gamma)$ by appealing to a new non-asymptotic upper bound on the average probability of error $P_e(\Phi_n; \Gamma)$.

III. REVIEW OF EXISTING RESULTS

In this section, we review existing asymptotic and non-asymptotic results for the three problems in Section II.

A. Existing Results for the WAK Problem

We first state the asymptotic optimal rate region for the WAK problem. Subsequently, we state some existing non-asymptotic bounds on the probability of error, defined in (8).

Let $\mathcal{P}(P_{XY})$ be the set of all joint distributions $P_{UXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$ such that the $\mathcal{X} \times \mathcal{Y}$ -marginal of P_{UXY} is the source distribution P_{XY} , $U - Y - X$ forms a Markov chain in that order and² $|\mathcal{U}| \leq |\mathcal{Y}| + 1$. Define

$$\mathcal{R}_{\text{WAK}}^* := \bigcup_{P_{UXY} \in \mathcal{P}(P_{XY})} \{ (R_1, R_2) \in \mathbb{R}_+^2 : R_1 \geq H(X|U), R_2 \geq I(U; Y) \}. \quad (37)$$

Wyner [2] and Ahlswede-Körner [3] proved the following:

Theorem 2 (Wyner [2], Ahlswede-Körner [3]). *For every $0 < \varepsilon < 1$, we have*

$$\mathcal{R}_{\text{WAK}}(\varepsilon) = \mathcal{R}_{\text{WAK}} = \mathcal{R}_{\text{WAK}}^*, \quad (38)$$

where $\mathcal{R}_{\text{WAK}}(\varepsilon)$ and \mathcal{R}_{WAK} are defined in (12) and (13) respectively.

To prove the direct part, Wyner used the PBL and the Markov lemma [2] while Ahlswede-Körner [3] used a maximal code construction. Only weak converses were provided in [2] and [3]. Ahlswede-Gács-Körner [31] proved the strong converse using entropy and image-size characterizations [30, Ch. 15], which are based on the so-called blowing-up lemma [30, Ch. 5]. See [30, Thm. 16.4].

²The cardinality bound on \mathcal{U} in the definition of $\mathcal{P}(P_{XY})$ is only applied when we consider the single letter characterization $\mathcal{R}_{\text{WAK}}^*$, and it is not applied when we consider the non-asymptotic analysis. Similar remarks are also applied for the WZ and GP problems.

Theorem 3 presents a non-asymptotic version of Wyner's bound and was proved recently by Kuzuoka [32] using the PBL technique and the Markov lemma [2]. For fixed auxiliary alphabet \mathcal{U} , joint distribution $P_{UXY} \in \mathcal{P}(P_{XY})$ and arbitrary non-negative constants γ_b and γ_c , we define two sets

$$\mathcal{T}_b^{\text{WAK}}(\gamma_b) := \left\{ (u, x) \in \mathcal{U} \times \mathcal{X} : \log \frac{1}{P_{X|U}(x|u)} \leq \gamma_b \right\}, \quad (39)$$

$$\mathcal{T}_c^{\text{WAK}}(\gamma_c) := \left\{ (u, y) \in \mathcal{U} \times \mathcal{Y} : \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \leq \gamma_c \right\}. \quad (40)$$

These sets are similar to the *typical* sets used extensively in network information theory [1] but note that these sets only involve the entropy and information densities. Consequently, the probabilities of these sets (events) are entropy and information spectrum quantities [7]. The subscripts b and c refer respectively to *binning* and *covering*. Similar subscripts and will be used in the sequel for the other side-information problems to demonstrate the similarities between the proof techniques all of which leverage on ideas from channel resolvability [7, Ch. 6] [14] and channel simulation [15]–[19].

Theorem 3 (Kuzuoka [32]). *For arbitrary $\gamma_b, \gamma_c \geq 0$, there exists a WAK code Φ with error probability satisfying*

$$P_e(\Phi) \leq 2\sqrt{P_{UX}(\mathcal{T}_b^{\text{WAK}}(\gamma_b)^c) + P_{UY}(\mathcal{T}_c^{\text{WAK}}(\gamma_c)^c)} + \frac{2^{\gamma_b}}{|\mathcal{M}|} + \exp \left\{ -\frac{|\mathcal{L}|}{2^{\gamma_c}} \right\}. \quad (41)$$

The first and second terms are the dominant ones for an appropriate choice of γ_b and γ_c . They are entropy and information spectrum quantities that can be easily evaluated in the n -fold i.i.d. setting using, for example, the central limit theorem. Observe that the second term represents the encoding of Y with U and the first term represents the decoding of X given U . The first term can be large due to the square root resulting from Wyner's PBL. Verdú [6] demonstrated a refined version of Theorem 3 in which the square root in the first term is removed.

Theorem 4 (Verdú [6]). *For arbitrary $\gamma_b, \gamma_c \geq 0$, there exists a WAK code Φ with error probability satisfying*

$$P_e(\Phi) \leq P_{UX}(\mathcal{T}_b^{\text{WAK}}(\gamma_b)^c) + P_{UY}(\mathcal{T}_c^{\text{WAK}}(\gamma_c)^c) + \frac{2^{\gamma_b}}{|\mathcal{M}|} + \exp \left\{ -\frac{|\mathcal{L}|}{2^{\gamma_c}} \right\}. \quad (42)$$

It is worth briefly describing the side-information encoder and the decoder used in Verdú's work [6]. Verdú defines the following metric

$$\pi(u, y) := \Pr \left\{ \log \frac{1}{P_{X|U}(X|u)} > \gamma_b \mid Y = y \right\} \quad (43)$$

and the helper encoder searches for a codeword $u_l, l \in \mathcal{L}$ that minimizes $\pi(u_l, y)$ given the side-information y . The decoder examines the appropriate decompression bin for a codeword with entropy density $-\log P_{X|U}(X|u)$ not exceeding γ_b . Verdú then uses generalizations of the covering and packing lemmas in [1] to prove (42). In Section IV-A, we further improve on Verdú's bound. We show that the two information spectrum terms in (42) (first two terms) can be combined under a single probability. Thus, the derived achievable second-order coding rate is better than that using Verdú's bound in (42).

In another line of work, Kelly and Wagner [33] demonstrated bounds on the error exponent for the WAK problem with stationary and memoryless source $P_{X^n Y^n} = P_{XY}^n$. Here we present only the direct part (lower bound on the error exponent).

Theorem 5 (Kelly-Wagner [33]). *There exists a sequence of WAK codes $\{\Phi_n\}_{n=1}^\infty$ with rates satisfying*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log |\mathcal{M}_n| \leq R_1, \quad \limsup_{n \rightarrow \infty} \frac{1}{n} \log |\mathcal{L}_n| \leq R_2, \quad (44)$$

such that the error probabilities satisfy

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_e(\Phi_n)} \geq \eta_L(P_{XY}, R_1, R_2) \quad (45)$$

where

$$\begin{aligned} \eta_L(P_{XY}, R_1, R_2) &:= \min_{Q_Y} \max_{Q_{U|Y}} \min_{\substack{Q_{X|YU}: \\ H(Q_X) \geq R_1}} D(Q_{XYU} || P_{XY} \times Q_{U|Y}) \\ &+ \begin{cases} |R_1 + R_2 - H(Q_{X|U}|Q_U) - I(Q_Y, Q_{U|Y})|^+ & I(Q_Y, Q_{U|Y}) \geq R_2 \\ |R_1 - H(Q_{X|U}|Q_U)|^+ & I(Q_Y, Q_{U|Y}) < R_2 \end{cases} \end{aligned} \quad (46)$$

The proof in [33] is based on the method of types [30]. The helper encoder quantizes its observation Y^n using the test channel $Q_{U|Y}$. It sends the quantization index and the type of Y^n . Then, the helper encoder optionally uses binning for the quantized sequence. The primary encoder uses binning for each source type class if necessary. This corresponds to the two cases in (46). The decoder finds the sequence in the specified bin with the specified type with the smallest empirical conditional entropy (conditioned on the side quantized sequence). Thus, the code is a universally attainable one.

B. Existing Results for the WZ Problem

As in the previous section, we state the asymptotic WZ rate-distortion function. Subsequently, we state some existing non-asymptotic bounds on the probability of excess distortion defined in (14).

Let $\mathcal{P}_D(P_{XY})$ be the set of all pairs (P_{UXY}, g) where $P_{UXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$ is a joint distribution and $g : \mathcal{U} \times \mathcal{Y} \rightarrow \mathcal{X}$ is a (reproduction) function such that the $\mathcal{X} \times \mathcal{Y}$ -marginal of P_{UXY} is the source distribution P_{XY} , $U - X - Y$ forms a Markov chain in that order, $|\mathcal{U}| \leq |\mathcal{X}| + 1$ and the distortion constraint is satisfied, i.e.,

$$\mathbb{E}[d(X, g(U, Y))] = \sum_{u,x,y} P_{UXY}(u, x, y) d(x, g(u, y)) \leq D. \quad (47)$$

Define the function

$$R_{WZ}^*(D) := \min_{(P_{UXY}, g) \in \mathcal{P}_D(P_{XY})} I(U; X) - I(U; Y). \quad (48)$$

Note from Markovity that $I(U; X) - I(U; Y) = I(U; X|Y)$. Then, we have the following asymptotic characterization of the WZ rate-distortion function.

Theorem 6 (Wyner-Ziv [4]). *We have*

$$R_{WZ}(D) = R_{WZ}^*(D), \quad (49)$$

where $R_{WZ}(D)$ is defined in (22).

The direct part of the proof of the theorem in the original Wyner-Ziv paper [4] is based on the average fidelity criterion in (23). It relies on the *compress-bin* idea. That is, binning is used to reduce the rate of the description of the main source to the receiver. The encoder transmits the bin index and the decoder searches within that bin for the transmitted codeword. The reproduction function g is then used to reproduce the source to within a distortion D . To prove Theorem 6 for the probability of excess distortion criterion, we can use the bounds provided in the following (e.g., Theorems 7 or 8), which, at a high-level, are based on the same ideas as in [4]. The converse in [4] was proved for the average fidelity criterion in (23) but can be adapted for the probability of excess distortion criterion by noting that for a sequence of rate- R codes³ $\{\Phi_n\}_{n=1}^\infty$ for which $P_e(\Phi_n; D) \rightarrow 0$,

$$\begin{aligned} &\limsup_{n \rightarrow \infty} \mathbb{E}[d_n(X^n, \psi_n(f_n(X^n), Y^n))] \\ &= \limsup_{n \rightarrow \infty} \int_0^{D_{\max}} \mathbb{P}(d_n(X^n, \psi_n(f_n(X^n), Y^n)) \geq t) dt \end{aligned} \quad (50)$$

$$\leq \int_0^{D_{\max}} \limsup_{n \rightarrow \infty} \mathbb{P}(d_n(X^n, \psi_n(f_n(X^n), Y^n)) \geq t) dt \quad (51)$$

$$= \int_0^D \limsup_{n \rightarrow \infty} \mathbb{P}(d_n(X^n, \psi_n(f_n(X^n), Y^n)) \geq t) dt \leq D \quad (52)$$

³This means that the \limsup of $R(\Phi_n)$ defined in (15) is no larger than R . Also see (58).

where (51) is by Fatou's lemma (the space $[0, D_{\max}]$ has finite Lebesgue measure) and the equality in (52) is by the assumption that $P_e(\Phi_n; t) \rightarrow 0$ for every $t \in (D, D_{\max}]$. As such by standard manipulations (see [1, Thm. 11.3]), $R_{\text{WZ}}(D) \geq R_{\text{WZ}}^*(D)$ for the probability of excess distortion criterion. To the best of the authors' knowledge, it is not known whether $R_{\text{WZ}}(\varepsilon, D) = R_{\text{WZ}}^*(D)$ for all $0 < \varepsilon < 1$, i.e., whether the strong converse holds for the probability of excess distortion criterion.

The analogue of Theorem 3 can be distilled from the work of Iwata and Muramatsu [9] who derived the general formula for the (fixed-length, maximum-distortion criterion) Wyner-Ziv problem. We rederive their general formula in Section V-B. Before we state the theorem, let us define the three sets for fixed $(P_{UXY}, g) \in \mathcal{P}(P_{XY})$ and non-negative constants γ_p and γ_c :

$$\mathcal{T}_p^{\text{WZ}}(\gamma_p) := \left\{ (u, y) \in \mathcal{U} \times \mathcal{Y} : \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \geq \gamma_p \right\} \quad (53)$$

$$\mathcal{T}_c^{\text{WZ}}(\gamma_c) := \left\{ (u, x) \in \mathcal{U} \times \mathcal{X} : \log \frac{P_{X|U}(y|u)}{P_X(x)} \leq \gamma_c \right\} \quad (54)$$

$$\mathcal{T}_d^{\text{WZ}}(D) := \{(u, x, y) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y} : d(x, g(u, y)) \leq D\}. \quad (55)$$

These sets have intuitive explanations: $\mathcal{T}_c^{\text{WZ}}(\gamma_c)^c$ represents the *covering* error that U is unable to describe X to the desired level indicated by γ_c ; $\mathcal{T}_p^{\text{WZ}}(\gamma_p)^c$ represents the *packing* error in which the decoder is unable to decode the correct codeword U given Y using a threshold test based on the information density statistic and γ_p ; $\mathcal{T}_d^{\text{WZ}}(D)^c$ represents the *distortion* error in which the reproduction \hat{X} not within a distortion of D of the source X .

Theorem 7 (Refinement of Iwata-Muramatsu [9]). *For arbitrary $\gamma_p, \gamma_c \geq 0$ and an arbitrary positive integer L , there exists a WZ code Φ with probability of excess distortion satisfying*

$$P_e(\Phi; D) \leq 2\sqrt{P_{UY}(\mathcal{T}_p^{\text{WZ}}(\gamma_p)^c)} + P_{UX}(\mathcal{T}_c^{\text{WZ}}(\gamma_c)^c) + P_{UXY}(\mathcal{T}_d^{\text{WZ}}(D)^c) + \frac{L}{2^{\gamma_p}|\mathcal{M}|} + \exp\left\{-\frac{L}{2^{\gamma_c}}\right\}. \quad (56)$$

The quantity $L/|\mathcal{M}|$ in (56) can be interpreted as the size of each bin. Recall that in WZ coding [4], binning is used to reduce the rate used to describe the source to the decoder, because the decoder has access to correlated side-information Y . Verdú improved on this bound by using a novel decoder and non-asymptotic versions of the packing and covering lemmas in [1]. This removes the factor 2 and the square-root operation in the first term in (56). These operations, which loosen the bound in (56), result from the use of Wyner's PBL and Markov lemma.

Theorem 8 (Verdú [6]). *For arbitrary $\gamma_p, \gamma_c \geq 0$ and an arbitrary positive integer L , there exists a WZ code Φ with probability of excess distortion satisfying*

$$P_e(\Phi; D) \leq P_{UY}(\mathcal{T}_p^{\text{WZ}}(\gamma_p)^c) + P_{UX}(\mathcal{T}_c^{\text{WZ}}(\gamma_c)^c) + P_{UXY}(\mathcal{T}_d^{\text{WZ}}(D)^c) + \frac{L}{2^{\gamma_p}|\mathcal{M}|} + \exp\left\{-\frac{L}{2^{\gamma_c}}\right\}. \quad (57)$$

The salient terms in (57) are the first three. The first two terms are information spectrum terms which can be evaluated easily in i.i.d. setting. The third term, though not an information spectrum term can also be bounded easily. In Section IV-B, we improve on Verdú's bound in (57) by placing *all three* salient terms under a single probability. Thus, the second-order coding rate derived from our bound is no worse than that derived from (57). In fact, one of our contributions is to show that the dispersion (or second-order coding rate) of lossy source coding derived by Ingber-Kochman [27] and Kostina-Verdú [28] can be derived from our non-asymptotic channel-simulation-type bound. See Theorem 29 in Section VI-B.

Kelly-Wagner also derived bounds on the error exponent for the probability of excess distortion in the WZ problem with stationary and memoryless source $P_{X^n Y^n} = P_{XY}^n$. The achievability result can be stated as follows:

Theorem 9 (Kelly-Wagner [33]). *For a fixed sequence of WZ codes $\{\Phi_n\}_{n=1}^\infty$, with rates*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log |\mathcal{M}_n| \leq R \quad (58)$$

let the error exponent for the probabilities of excess distortion of $\{\Phi_n\}_{n=1}^\infty$ be defined as

$$\theta(R, D, P_{XY}) := \liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_e(\Phi_n)} \quad (59)$$

There exists a sequence of WZ codes $\{\Phi_n\}_{n=1}^\infty$ with rates satisfying (58) for which the error exponent in (59) is lower bounded as

$$\theta(R, D, P_{XY}) \geq \min_{Q_X} \max_{Q_{U|X}} \min_{Q_Y} \max_{g \in \mathcal{G}} \min_{Q_{UXY}} J_D(Q_{UXY}, P_{XY}, f, R, D) \quad (60)$$

where the set $\mathcal{G} := \{g : \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}\}$, the auxiliary alphabet \mathcal{U} takes on finitely many values and

$$J_D(Q_{UXY}, P_{XY}, g, R, D) := \begin{cases} D(Q_{UXY} \| P_{XY} \times Q_{U|X}) & \mathbb{E}_{Q_{UXY}}[d(X, g(U, Y))] \geq D \\ D(Q_{UXY} \| P_{XY} \times Q_{U|X}) \\ \quad + |R - I(Q_X, Q_{U|X})| & \mathbb{E}_{Q_{UXY}}[d(X, g(U, Y))] < D, \\ \quad + I(Q_Y, Q_{U|Y})^+ & I(Q_X, Q_{U|X}) \geq R \\ \infty & \text{otherwise} \end{cases} \quad (61)$$

Note in the final minimization over Q_{UXY} , $Q_{U|X}$ and Q_Y are fixed to be those specified earlier in the optimization.

The proof of Theorem 9 is similar to Theorem 5 and makes heavy use of the method of types. Roughly speaking, the two cases in (61) correspond to the whether binning is necessary based on the realized source type $Q_X \in \mathcal{P}_n(\mathcal{X})$. Also, for every source type Q_X , we can optimize over the test channel (shell) $Q_{U|X} \in \mathcal{V}_n(\mathcal{U}; Q_X)$. For every side-information type $Q_Y \in \mathcal{P}_n(\mathcal{Y})$, we can optimize over the reproduction function g . This explains the first four optimizations in (60). The final minimization represents an adversarial consistent joint distribution for the triple $(U, X, Y) \sim Q_{UXY}$. This result, just like the WAK one in Theorem 5 has a game-theoretic interpretation and the order of plays matches the problem.

C. Existing Results for the GP Problem

We conclude this section by stating existing results for the GP problem [5]. Recall that in the GP problem, we have a channel $W : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{Y}$ and a state distribution $P_S \in \mathcal{P}(\mathcal{S})$. Assume for simplicity that all alphabets are finite sets. Let $\mathcal{P}_\Gamma(W, P_S)$ be the collection of all joint distributions $P_{UXSY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y})$ such that the \mathcal{S} -marginal is P_S , the conditional distribution $P_{Y|XS} = W$, $U - (X, S) - Y$ forms a Markov chain in that order,

$$\mathbb{E}[g(X)] \leq \Gamma \quad (62)$$

and $|\mathcal{U}| \leq \min\{|\mathcal{X}||\mathcal{S}|, |\mathcal{S}| + |\mathcal{Y}| - 1\}$. Define the quantity

$$C_{\text{GP}}^*(\Gamma) := \max_{P_{UXSY} \in \mathcal{P}_\Gamma(W, P_S)} I(U; Y) - I(U; S), \quad (63)$$

where $I(U; Y)$ and $I(U; S)$ are computed with respect to the joint distribution P_{UXSY} . If there is no cost constraint (62), we simply write C_{GP}^* instead of $C_{\text{GP}}^*(\infty)$. Then, we have the following asymptotic characterization.

Theorem 10 (Gel'fand-Pinsker [5]). *If the alphabets \mathcal{S}, \mathcal{X} and \mathcal{Y} are discrete, for every $0 < \varepsilon < 1$, we have*

$$C_{\text{GP}}(\varepsilon) = C_{\text{GP}} = C_{\text{GP}}^* \quad (64)$$

where $C_{\text{GP}}(\varepsilon)$ and C_{GP} are defined in (35) and (36) respectively.

The direct part was proved using a covering-packing argument as well as the conditional typicality lemma (using the notion of strong typicality). Essentially, each message $m \in \mathcal{M}$ is uniquely associated to a subcodebook of size L . To send message m , the encoder looks in the m -th subcodebook for a codeword that is jointly typical with the noncausal state. The decoder then searches for the unique subcodebook which contains at least one codeword that is jointly typical with the channel output. The weak converse in the original Gel'fand-Pinsker paper was proved using the Csiszár-sum-identity. See [1, Thm. 7.3]. Tyagi and Narayan proved a strong converse [36] using entropy and image-size characterizations via judicious choices of auxiliary channels. Their proof only applies to discrete memoryless channels with discrete state distribution without cost constraints.

A recent non-asymptotic bound for the average error probability was proved by Tan [10] without the cost constraint (24). To state the bound, we define the sets

$$\mathcal{T}_p^{\text{GP}}(\gamma_p) := \left\{ (u, y) \in \mathcal{U} \times \mathcal{Y} : \log \frac{P_{Y|U}(Y|U)}{P_Y(Y)} \geq \gamma_p \right\} \quad (65)$$

$$\mathcal{T}_c^{\text{GP}}(\gamma_c) := \left\{ (u, s) \in \mathcal{U} \times \mathcal{S} : \log \frac{P_{S|U}(S|U)}{P_S(S)} \leq \gamma_c \right\} \quad (66)$$

These are analogous to the typical sets used extensively in network information theory [1] but they only involve the information densities. The first set in (65) represents *packing* event while the second in (66) represents *covering* event. Also notice the similarities to $\mathcal{T}_p^{\text{WZ}}(\gamma_p)$ and $\mathcal{T}_c^{\text{WZ}}(\gamma_c)$ in (53) and (54) for the WZ problem. However, now S plays the role of X while Y has the same interpretation in both the WZ and GP problems. With these definitions, Tan [10] proved the following non-asymptotic characterization for the average error probability in the GP problem.

Theorem 11 (Tan [10]). *For arbitrary $\gamma_p, \gamma_c \geq 0$ and an arbitrary positive integer L , there exists a GP code Φ with average probability of error satisfying*

$$P_e(\Phi) \leq 2\sqrt{P_{UY}(\mathcal{T}_p^{\text{GP}}(\gamma_p)^c) + P_{US}(\mathcal{T}_c^{\text{GP}}(\gamma_c)^c)} + \frac{|\mathcal{M}|L}{2^{\gamma_p}} + \exp\left\{-\frac{L}{2^{\gamma_c}}\right\}. \quad (67)$$

The parameter L again represents the number of codewords in each subcodebook. The first two terms are information spectrum terms which can be evaluated fairly easily in the n -shot setting. The proof of (67) again uses Wyner's PBL and the Markov lemma [2]. Verdú [6] presented a tightened version of the above bound for the case where there are no cost constraints.

Theorem 12 (Verdú [6]). *For arbitrary $\gamma_p, \gamma_c \geq 0$ and an arbitrary positive integer L , there exists a GP code Φ with average probability of error satisfying*

$$P_e(\Phi) \leq P_{UY}(\mathcal{T}_p^{\text{GP}}(\gamma_p)^c) + P_{US}(\mathcal{T}_c^{\text{GP}}(\gamma_c)^c) + \frac{|\mathcal{M}|L}{2^{\gamma_p}} + \exp\left\{-\frac{L}{2^{\gamma_c}}\right\}. \quad (68)$$

To prove (68), Verdú considered the function

$$\zeta(s, u) := \Pr\left\{\log \frac{P_{Y|U}(Y|U)}{P_Y(Y)} < \gamma_p \mid (U, S) = (u, s)\right\} \quad (69)$$

To send message m , the encoder given the noncausal state sequence s searches within the m -th subcodebook for the codeword u_{ml} that minimizes $\zeta(s, u_{ml})$. The decoder is Feinstein-like [50] and basically declares finds a codeword whose information density $\log \frac{P_{Y|U}(Y|U)}{P_Y(Y)}$ exceeds a prescribed threshold γ_p . Its subcodebook index is declared as the sent message. In Section IV-C, by using a technique based on channel-simulation, we show that the information spectrum terms in (68) can, in fact, be placed under a single probability. Thus, the derived achievable second-order coding rate is better than that using Verdú's bound in (68).

When the channel and state are discrete, memoryless and stationary, i.e., $W^n(y^n|x^n, s^n) = \prod_{i=1}^n W(y_i|x_i, s_i)$ and $P_{S^n}(s^n) = \prod_{i=1}^n P_S(s_i)$, error exponents can be derived. Indeed, a random coding error exponent for the GP problem was presented by Moulin and Wang [35] for the case without cost constraints. We state a simplified version of the main result in [35] here.

Theorem 13 (Moulin-Wang [35]). *There exists a sequence of GP codes $\{\Phi_n\}_{n=1}^\infty$ of rates*

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log |\mathcal{M}_n| \geq R \quad (70)$$

for which the average error probabilities satisfy

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_e(\Phi_n)} \geq \eta_L(W, P_S, R) \quad (71)$$

where, letting the auxiliary alphabet \mathcal{U} take on finitely many values, we have

$$\begin{aligned} \eta_L(W, P_S, R) := \min_{Q_S} \max_{Q_{XU|S}} \min_{Q_{Y|XUS}} & \left\{ D(Q_S \times Q_{XU|S} \times Q_{Y|XUS} \| P_S \times Q_{XU|S} \times W) \right. \\ & \left. + |I_Q(U; Y) - I_Q(U; S) - R|^+ \right\}. \end{aligned} \quad (72)$$

Note that $I_Q(U; Y)$ and $I_Q(U; S)$ are the mutual information quantities computed with respect to $Q_{SXUY} := Q_S \times Q_{XU|S} \times Q_{Y|XUS}$.

Like Theorem 9, this result has a game-theoretic interpretation. Nature chooses an adversarial state sequence represented by the type $Q_S \in \mathcal{P}_n(\mathcal{S})$ and the player can optimize for the best test channel $Q_{XU|S} \in \mathcal{V}_n(\mathcal{X} \times \mathcal{U}; Q_S)$ to find a good channel input. Nature then chooses an adversarial channel $Q_{Y|XUS} \in \mathcal{V}_n(\mathcal{Y}; Q_{XU|S} Q_S)$. The order

of plays matches the order of the optimizations in (72). It is worth mentioning that a sphere-packing bound (upper bound on the error exponent) was derived by Tyagi and Narayan [36] by appealing to their strong converse and the Haroutunian change of measure technique [51].

IV. MAIN RESULTS: NOVEL NON-ASYMPTOTIC ACHIEVABILITY BOUNDS

In this section, we describe our results concerning novel non-asymptotic achievability bounds for the WAK, WZ and GP problems. We show using ideas from channel resolvability [7, Ch. 6] [13] [14] and channel simulation [15]–[19] that the bounds obtained in Theorems 4, 8 and 12 can be refined so as to obtain better second-order coding rates. The definition of and techniques involving channel resolvability and channel simulation are reviewed in Appendices B and C respectively. These are concepts that form crucial components of the proofs of the Channel-Simulation-type (CS-type) bounds in the sequel.

The following quantity, introduced in [14], will be used extensively in this section so we provide its definition here. For a joint distribution $P_{UY} \in \mathcal{P}(\mathcal{U} \times \mathcal{Y})$ and a positive constant γ_c , define

$$\Delta(\gamma_c, P_{UY}) := \sum_{(u,y) \in \mathcal{U} \times \mathcal{Y}} \frac{P_U(u)P_{Y|U}(y|u)^2}{P_Y(y)} \mathbf{1} \left\{ \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \leq \gamma_c \right\} \quad (73)$$

$$= \mathbb{E}_{P_{UY}} \left[\frac{P_{Y|U}(Y|U)}{P_Y(Y)} \mathbf{1} \left\{ \frac{P_{Y|U}(Y|U)}{P_Y(Y)} \leq 2^{\gamma_c} \right\} \right]. \quad (74)$$

Observe from (74) that $\Delta(\gamma_c, P_{UY})$ has the property that

$$\Delta(\gamma_c, P_{UY}) \leq 2^{\gamma_c}. \quad (75)$$

A. Novel Non-Asymptotic Achievability Bound for the WAK Problem

Fix an auxiliary alphabet \mathcal{U} and a joint distribution $P_{UXY} \in \mathcal{P}(P_{XY})$. See definition of $\mathcal{P}(P_{XY})$ prior to (37). Also recall the definitions of the sets $\mathcal{T}_b^{\text{WAK}}(\gamma_b)$ and $\mathcal{T}_c^{\text{WAK}}(\gamma_c)$ for the WAK problem in (39) and (40) respectively. The following is our CS-type bound.

Theorem 14 (CS-type bound for WAK coding). *For arbitrary $\gamma_b, \gamma_c \geq 0$, and $\delta > 0$, there exists a WAK code Φ with error probability satisfying*

$$\begin{aligned} P_e(\Phi) &\leq P_{UXY} [(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c] \\ &\quad + \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \sqrt{\frac{\Delta(\gamma_c, P_{UY})}{|\mathcal{L}|}} + \delta. \end{aligned} \quad (76)$$

See Appendix D for the proof of Theorem 14. Observe that the primary novelty of the bound in (76) lies in the fact that both error events $\{(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c\}$ and $\{(u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c\}$ lie under the same probability and so can be bounded together (as a vector) in second-order coding analysis. See Section VI-A. Notice that the sum of the information spectrum terms (first two terms) in Verdú's bound in (42) is the result upon invoking the union bound on the first term in (76). We illustrate the differences in the resulting second-order coding rates numerically in Section VII. The bound in (76) is rather unwieldy. We can simplify it without losing too much. Indeed, using the definition of $\mathcal{T}_b^{\text{WAK}}(\gamma_b)$, we observe that the third term in (76) can be bounded as

$$\frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) = \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) \frac{P_{X|U}(\tilde{x}|u)}{P_{X|U}(\tilde{x}|u)} \quad (77)$$

$$\leq \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) P_{X|U}(\tilde{x}|u) 2^{\gamma_b} \quad (78)$$

$$\leq \frac{2^{\gamma_b}}{|\mathcal{M}|}. \quad (79)$$

Together with (75), we have the following simplified CS-type bound, which resembles a Feinstein-type [50] achievability bound.

Corollary 15 (Simplified CS-type bound for WAK coding). *For arbitrary $\gamma_b, \gamma_c \geq 0$, and $\delta > 0$, there exists a WAK code Φ with error probability satisfying*

$$P_e(\Phi) \leq P_{UXY} [(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c] + \frac{2^{\gamma_b}}{|\mathcal{M}|} + \sqrt{\frac{2^{\gamma_c}}{|\mathcal{L}|}} + \delta. \quad (80)$$

If (X^n, Y^n) is drawn from the product distribution P_{XY}^n , then by designing γ_b and γ_c appropriately, we see that the dominating term in (80) is the first one. The other terms vanish with n . In particular, δ stems from the amount of common randomness known to all parties and since the amount of common randomness can be arbitrarily large, δ can be made arbitrarily small. In addition, Δ in (73) results from approximating an arbitrary distribution with one that is simulated by a channel [14, Lem. 2]. See Appendix B.

By modifying the helper in the proof of Theorem 14, we can show the following theorem.

Theorem 16 (Modified CS-type bound for WAK coding). *For arbitrary $\gamma_b, \gamma_c \geq 0$, and $\delta > 0$ and positive integer J , there exists a WAK code Φ with error probability satisfying*

$$P_e(\Phi) \leq P_{UXY} [(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c] + \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \frac{J}{|\mathcal{M}||\mathcal{L}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \sqrt{\frac{\Delta(\gamma_c, P_{UY})}{J}} + \delta. \quad (81)$$

See Appendix E for the proof of Theorem 16. By letting $J = |\mathcal{L}|$ in (81), we recover (76) up to an additional residual term, which is unimportant in second-order analysis. A close inspection of the proof reveals that the additional term is due to additional random bin coding at the helper, which is not needed if $J = |\mathcal{L}|$.

Remark 1. *For the special case such that test channel $P_{U|Y}$ is noiseless, we can show that there exists a WAK code satisfying*

$$P_e(\Phi) \leq P_{XY} [(x, y) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup \mathcal{T}_s^{\text{WAK}}(\gamma_s)^c] + \frac{2^{\gamma_b}}{|\mathcal{M}|} + \frac{2^{\gamma_s}}{|\mathcal{M}||\mathcal{L}|} \quad (82)$$

for any $\gamma_b, \gamma_s \geq 0$, where

$$\mathcal{T}_s^{\text{WAK}}(\gamma_s) := \left\{ (x, y) \in \mathcal{X} \times \mathcal{Y} : \log \frac{1}{P_{XY}(x, y)} \leq \gamma_s \right\}. \quad (83)$$

We can prove the bound (82) by using the standard Slepian-Wolf type bin coding for both the main encoder and the helper [25], [47]. As it will turn out later in Section VI-A, this simple bound gives tighter second-order achievability in some cases.

B. Novel Non-Asymptotic Achievability Bound for the WZ Problem

We now turn our attention to the WZ problem where we derive a similar bound as in Theorem 14. This improves on Verdú's bound in Theorem 8. It again uses the same CS idea for the covering part. Recall the definitions of the sets $\mathcal{T}_p^{\text{WZ}}(\gamma_p)$, $\mathcal{T}_c^{\text{WZ}}(\gamma_c)$ and $\mathcal{T}_d^{\text{WZ}}(D)$ in (53), (54) and (55) respectively. In the following, $P_{UXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$ satisfying (i) the $\mathcal{X} \times \mathcal{Y}$ -marginal of P_{UXY} is P_{XY} and (ii) $U - X - Y$ forms a Markov chain is fixed.

Theorem 17 (CS-type bound for WZ coding). *For arbitrary constants $\gamma_p, \gamma_c \geq 0$, and $\delta > 0$ and positive integer L , there exists a WZ code Φ with probability of excess distortion satisfying*

$$P_e(\Phi; D) \leq P_{UXY} [(u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)^c \cup (u, x) \in \mathcal{T}_c^{\text{WZ}}(\gamma_c)^c \cup (u, x, y) \in \mathcal{T}_d^{\text{WZ}}(D)^c] + \frac{L}{|\mathcal{M}|} \sum_{(u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)} P_U(u) P_Y(y) + \sqrt{\frac{\Delta(\gamma_c, P_{UX})}{L}} + \delta \quad (84)$$

where $\Delta(\gamma_c, P_{UX})$ is defined in (73).

The proof of Theorem 19 is provided in Appendix F. As with Theorem 14, the main novelty of our bound lies in the fact that the three error events lie under the same probability, making it amendable to treat all three error events *jointly*. The residual terms in (84) (namely, the second, third and fourth terms) are relatively small with a

proper choice of constants $\gamma_p, \gamma_c, \delta \geq 0$ and $L \in \mathbb{N}$ as we shall see in the sequel. We can again relax the somewhat cumbersome second and third terms in (84) by noting the definition of $\mathcal{T}_p^{\text{WZ}}(\gamma_p)$ and by going through the same steps to upper bound Δ ; cf. (75). We thus obtain:

Corollary 18 (Simplified CS-type bound for WZ coding). *For arbitrary constants $\gamma_p, \gamma_c \geq 0$, and $\delta > 0$ and positive integer L , there exists a WZ code Φ with probability of excess distortion satisfying*

$$P_e(\Phi; D) \leq P_{USXY} [(u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)^c \cup (u, x) \in \mathcal{T}_c^{\text{WZ}}(\gamma_c)^c \cup (u, x, y) \in \mathcal{T}_d^{\text{WZ}}(D)^c] \\ + \frac{L}{2^{\gamma_p} |\mathcal{M}|} + \sqrt{\frac{2^{\gamma_c}}{L}} + \delta. \quad (85)$$

To obtain achievable second-order coding rates for the WZ problem, we evaluate the bound in (85) for appropriate choices of $\gamma_p, \gamma_c \geq 0$, and $\delta > 0$ and $L \in \mathbb{N}$ in Section VI-B. Since the lossy source coding problem is a special case of WZ coding, we use a specialization of the bound in (85) to derive an achievable dispersion (or second-order coding rate) of lossy source coding [27], [28], which turns out to be tight.

C. Novel Non-Asymptotic Achievability Bound for the GP Problem

This section presents with a novel non-asymptotic achievability bound for the GP problem, which is the dual of the WZ problem [49]. Our bound improves on both Theorems 11 and 12 and uses the same Channel-Simulation idea for the covering part. Recall the definitions of the sets $\mathcal{T}_p^{\text{GP}}(\gamma_p)$, $\mathcal{T}_c^{\text{GP}}(\gamma_c)$ and $\mathcal{T}_g^{\text{GP}}(\Gamma)$ in (65), (66) and (26) respectively. In the following, $P_{USXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{S} \times \mathcal{X} \times \mathcal{Y})$ satisfying (i) the \mathcal{S} -marginal of P_{USXY} is P_S , (ii) $P_{Y|XS} = W$ and (iii) $U - (X, S) - Y$ forms a Markov chain is fixed.

Theorem 19 (CS-type bound for GP coding). *For arbitrary constants $\gamma_p, \gamma_c \geq 0$, and $\delta > 0$ and positive integer L , there exists a GP code Φ with average error probability satisfying*

$$P_e(\Phi; \Gamma) \leq P_{USXY} [(u, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)^c \cup (u, s) \in \mathcal{T}_c^{\text{GP}}(\gamma_c)^c \cup x \in \mathcal{T}_g^{\text{GP}}(\Gamma)^c] \\ + L|\mathcal{M}| \sum_{(u, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)} P_U(u) P_Y(y) + \sqrt{\frac{\Delta(\gamma_c, P_{US})}{L}} + \delta \quad (86)$$

where $\Delta(\gamma_c, P_{US})$ is defined in (73).

The proof of Theorem 19 is provided in Appendix G. Notice that unlike the existing asymptotic and non-asymptotic results for GP coding (Theorems 11, 12 and 13), the channel input x satisfies the cost constraint (24) or its almost sure equivalent (cf. Proposition 1). Direct application of (75) to bound $\Delta(\gamma_c, P_{US})$ and the definition of $\mathcal{T}_p^{\text{GP}}(\gamma_p)$ in (65) yields the following:

Corollary 20 (Simplified CS-type bound for GP coding). *For arbitrary constants $\gamma_p, \gamma_c \geq 0$, and $\delta > 0$ and positive integer L , there exists a GP code Φ with average error probability satisfying*

$$P_e(\Phi; \Gamma) \leq P_{USXY} [(u, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)^c \cup (u, s) \in \mathcal{T}_c^{\text{GP}}(\gamma_c)^c \cup x \in \mathcal{T}_g^{\text{GP}}(\Gamma)^c] + \frac{L|\mathcal{M}|}{2^{\gamma_p}} + \sqrt{\frac{2^{\gamma_c}}{L}} + \delta. \quad (87)$$

To obtain achievable second-order coding rates for the GP problem, we evaluate the bound in (87) for appropriate choices of $\gamma_p, \gamma_c, \delta \geq 0$ and $L \in \mathbb{N}$ in Section VI-C.

V. GENERAL FORMULAS

In this section, we use the simplified CS-type bounds in Corollaries 15, 18 and 20 to derive achievable general formulas for the optimal rate region of the WAK problem, the rate-distortion function of the WZ problem and the capacity of the GP problem. This allows us to recover known results in [8]–[10]. By *general formula*, we mean that we consider sequences of these problems and do not place any underlying structure such as stationarity, memorylessness and ergodicity on the source and channel [7], [22]. To state our results, let us first recall the following probabilistic limit operations. Their properties are similar to the limit superior and limit inferior for numerical sequences in mathematical analysis and are summarized in [7].

Definition 7. Let $\mathbf{U} := \{U_n\}_{n=1}^\infty$ be a sequence of real-valued random variables. The limit superior in probability of \mathbf{U} is defined as

$$\mathbf{p}\text{-}\limsup_{n \rightarrow \infty} U_n := \inf \left\{ \alpha \in \mathbb{R} : \lim_{n \rightarrow \infty} \Pr(U_n > \alpha) = 0 \right\}. \quad (88)$$

The limit inferior in probability of \mathbf{U} is defined as

$$\mathbf{p}\text{-}\liminf_{n \rightarrow \infty} U_n := -\mathbf{p}\text{-}\limsup_{n \rightarrow \infty} (-U_n) \quad (89)$$

We also recall the following definitions from Han [7]. These definitions play a prominent role in the rest of this section.

Definition 8. Given a pair of stochastic processes $(\mathbf{X}, \mathbf{Y}) = \{X^n, Y^n\}_{n=1}^\infty$ with joint distributions $\{P_{X^n, Y^n}\}_{n=1}^\infty$, the spectral sup-mutual information rate is defined as

$$\bar{I}(\mathbf{X}; \mathbf{Y}) := \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)}. \quad (90)$$

The spectral inf-mutual information rate $\underline{I}(\mathbf{X}; \mathbf{Y})$ is defined as in (88) with $\mathbf{p}\text{-}\liminf$ in place of $\mathbf{p}\text{-}\limsup$. The spectral sup- and inf-conditional mutual information rates are defined similarly.

The spectral sup-conditional entropy rates is defined as

$$\bar{H}(\mathbf{Y}|\mathbf{X}) := \mathbf{p}\text{-}\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_{Y^n|X^n}(Y^n|X^n)}. \quad (91)$$

The spectral inf-conditional entropy rates is defined as in (91) with $\mathbf{p}\text{-}\liminf$ in place of $\mathbf{p}\text{-}\limsup$.

A. General Formula for the WAK problem

In this section, we consider sequences of the WAK problem indexed by the blocklength n where the sequence of source distributions $\{P_{X^n Y^n}\}_{n=1}^\infty$ is *general*, i.e., we do not place any assumptions on the structure of the source such as stationarity, memorylessness and ergodicity. We aim to characterize an inner bound to the optimal rate region defined in (13). We show that our inner bound coincides with that derived by Miyake and Kanaya [8] but is derived based on the upper bound on the error probability provided in our CS-type bound in Corollary 15. The choice of the parameters γ_b, γ_c and δ plays a crucial role and guides our choice of these parameters for second-order coding analysis in the following section.

Let $\mathcal{P}(\{P_{X^n Y^n}\}_{n=1}^\infty)$ be the set of all sequences of distributions $\{P_{U^n X^n Y^n}\}_{n=1}^\infty$ such that for every $n \geq 1$, $U^n - Y^n - X^n$ forms a Markov chain and the $(\mathcal{X}^n \times \mathcal{Y}^n)$ -marginal of $P_{U^n X^n Y^n}$ is $P_{X^n Y^n}$. Define the set

$$\hat{\mathcal{R}}_{\text{WAK}}^* := \bigcup_{\{P_{U^n X^n Y^n}\}_{n=1}^\infty \in \mathcal{P}(\{P_{X^n Y^n}\}_{n=1}^\infty)} \left\{ (R_1, R_2) \in \mathbb{R}_+^2 : R_1 \geq \bar{H}(\mathbf{X}|\mathbf{U}), R_2 \geq \bar{I}(\mathbf{U}; \mathbf{Y}) \right\} \quad (92)$$

Theorem 21 (Inner Bound to the Optimal Rate Region for WAK [8]). *We have*

$$\hat{\mathcal{R}}_{\text{WAK}}^* \subset \mathcal{R}_{\text{WAK}}. \quad (93)$$

We remark that by using techniques from [37], Miyake and Kanaya [8] showed that (93) is in fact an equality, i.e., $\hat{\mathcal{R}}_{\text{WAK}}^*$ is also an outer bound to \mathcal{R}_{WAK} . In addition, when the source distributions $\{P_{X^n Y^n}\}_{n=1}^\infty$ are stationary and memoryless (and the alphabets \mathcal{X} and \mathcal{Y} are discrete and finite), $\hat{\mathcal{R}}_{\text{WAK}}^*$ reduces to the single-letter region $\mathcal{R}_{\text{WAK}}^*$ defined in (37). This follows easily from the law of large numbers. The proof of Theorem 21 follows directly from the finite blocklength bound in Corollary 15. In fact, the weaker bounds in Theorems 3 and 4 suffice for this purpose.

Proof: Consider (80) and let us fix a process $\{P_{U^n X^n Y^n}\}_{n=1}^\infty \in \mathcal{P}(\{P_{X^n Y^n}\}_{n=1}^\infty)$ and a constant $\eta > 0$. Set

$$\frac{1}{n} \log |\mathcal{M}| := \bar{H}(\mathbf{X}|\mathbf{U}) + 2\eta \quad (94)$$

$$\frac{1}{n} \log |\mathcal{L}| := \bar{I}(\mathbf{U}; \mathbf{Y}) + 2\eta \quad (95)$$

$$\gamma_b := n(\bar{H}(\mathbf{X}|\mathbf{U}) + \eta) \quad (96)$$

$$\gamma_c := n(\bar{I}(\mathbf{U}; \mathbf{Y}) + \eta) \quad (97)$$

$$\delta := 2^{-n} \quad (98)$$

Then for blocklength n , the probability on the RHS of (80) can be written as

$$P_{U^n X^n Y^n} \left[\left\{ \frac{1}{n} \log \frac{1}{P_{X^n|U^n}(X^n|U^n)} \geq \bar{H}(\mathbf{X}|\mathbf{U}) + \eta \right\} \cup \left\{ \frac{1}{n} \log \frac{P_{Y^n|U^n}(Y^n|U^n)}{P_{Y^n}(Y^n)} \geq \bar{I}(\mathbf{U}; \mathbf{Y}) + \eta \right\} \right] \quad (99)$$

By the definition of the spectral sup-entropy rate and the spectral sup-mutual information rate, the probabilities of both events in (99) tend to zero. Further,

$$\frac{2^{\gamma_b}}{|\mathcal{M}|} = 2^{-n\eta} \rightarrow 0, \quad \sqrt{\frac{2^{\gamma_c}}{|\mathcal{L}|}} = 2^{-n\eta/2} \rightarrow 0. \quad (100)$$

Hence, $P_e(\Phi_n) \rightarrow 0$. Since $\eta > 0$ is arbitrary, from (94) and (95) we deduce that any pair of rates (R_1, R_2) satisfying $R_1 > \bar{H}(\mathbf{X}|\mathbf{U})$ and $R_2 > \bar{I}(\mathbf{U}; \mathbf{Y})$ is achievable. ■

B. General Formula for the WZ problem

In a similar way, we can recover the general formula for WZ coding derived by Iwata and Muramatsu [9]. Note however, that we directly work with the probability of excess distortion, which is related to but different from the maximum-distortion criterion employed in [9]. Once again, we assume that the source is $\{P_{X^n Y^n}\}_{n=1}^\infty$ is *general* in the sense explained in Section V-A.

Let $\mathcal{P}_D(\{P_{X^n Y^n}\}_{n=1}^\infty)$ be the set of all sequences of distributions $\{P_{U^n X^n Y^n}\}_{n=1}^\infty$ and reproduction functions $\{g_n : \mathcal{U}^n \times \mathcal{Y}^n \rightarrow \hat{\mathcal{X}}^n\}$ such that for every $n \geq 1$, $U^n - X^n - Y^n$ forms a Markov chain, the $(\mathcal{X}^n \times \mathcal{Y}^n)$ -marginal of $P_{U^n X^n Y^n}$ is $P_{X^n Y^n}$ and

$$\mathbf{p}\text{-}\limsup_{n \rightarrow \infty} d_n(X^n, g_n(U^n, Y^n)) \leq D \quad (101)$$

Define the rate-distortion function

$$\hat{R}_{\text{WZ}}^*(D) := \inf \{ \bar{I}(\mathbf{U}; \mathbf{X}) - \underline{I}(\mathbf{U}; \mathbf{Y}) \} \quad (102)$$

where the infimum is over all $\{P_{U^n X^n Y^n}, g_n\}_{n=1}^\infty \in \mathcal{P}_D(\{P_{X^n Y^n}\}_{n=1}^\infty)$.

Theorem 22 (Upper Bound to the Rate-Distortion Function for WZ [9]). *We have*

$$R_{\text{WZ}}(D) \leq \hat{R}_{\text{WZ}}^*(D). \quad (103)$$

Iwata and Muramatsu [9] showed in fact that (103) is an equality by proving a converse along the lines of [37]. It can be shown that the general rate-distortion function defined in (102) reduces to the one derived by Wyner and Ziv [4] in the case where the alphabets are finite and the source is stationary and memoryless.

Proof: Let $\eta > 0$. We start from the bound on the probability of excess distortion in (85), where we first consider $D + \eta$ instead of D . Let us fix the sequence of distribution and the sequence of functions $\{(P_{U^n X^n Y^n}, g_n)\}_{n=1}^\infty \in \mathcal{P}_D(\{P_{X^n Y^n}\}_{n=1}^\infty)$. Set

$$\frac{1}{n} \log |\mathcal{M}| := \bar{I}(\mathbf{U}; \mathbf{X}) - \underline{I}(\mathbf{U}; \mathbf{Y}) + 4\eta \quad (104)$$

$$\frac{1}{n} \log L := \bar{I}(\mathbf{U}; \mathbf{X}) + 2\eta \quad (105)$$

$$\gamma_p := n(\underline{I}(\mathbf{U}; \mathbf{Y}) - \eta) \quad (106)$$

$$\gamma_c := n(\bar{I}(\mathbf{U}; \mathbf{X}) + \eta) \quad (107)$$

and δ as in (98). Then, the probability in (85) for blocklength n can be written as

$$P_{U^n X^n Y^n} \left[\left\{ \frac{1}{n} \log \frac{P_{Y^n|U^n}(Y^n|U^n)}{P_{Y^n}(Y^n)} \leq \underline{I}(\mathbf{U}; \mathbf{Y}) - \eta \right\} \cup \left\{ \frac{1}{n} \log \frac{P_{X^n|U^n}(X^n|U^n)}{P_{X^n}(X^n)} \geq \bar{I}(\mathbf{U}; \mathbf{X}) + \eta \right\} \cup \left\{ d_n(X^n, g_n(U^n, Y^n)) \geq D + \eta \right\} \right] \quad (108)$$

By the definition of the spectral sup- and inf-mutual information rates and the distortion condition in (101), we observe that the probability in (108) tends to zero as n grows. By a similar calculation as in (100), the other terms in (85) also tend to zero. Hence, the probability of excess distortion $P_e(\Phi_n; D + \eta) \rightarrow 0$ as n grows. This holds for every $\eta > 0$. By (104), the any rate below $\bar{I}(\mathbf{U}; \mathbf{X}) - \underline{I}(\mathbf{U}; \mathbf{Y}) + 4\eta$ is achievable. In order to complete the proof, we choose a positive sequence satisfying $\eta_1 > \eta_2 > \dots > 0$ and $\eta_k \rightarrow 0$ as $k \rightarrow \infty$. Then, by using the *diagonal line argument* [7, Thm. 1.8.2], we complete the proof of (103). ■

C. General Formula for the GP problem

We conclude this section by showing that the non-asymptotic bound on the average probability of error derived in Corollary 20 can be adapted to recover the general formula for the GP problem derived in Tan [10]. Here, both the state distribution $\{P_{S^n} \in \mathcal{P}(\mathcal{S}^n)\}_{n=1}^\infty$ and the channel $\{W^n : \mathcal{X}^n \times \mathcal{S}^n \rightarrow \mathcal{Y}^n\}_{n=1}^\infty$ are general. In particular, the only requirement on the stochastic mapping W^n is that for every $(x^n, s^n) \in \mathcal{X}^n \times \mathcal{S}^n$,

$$\sum_{y^n \in \mathcal{Y}^n} W^n(y^n | x^n, s^n) = 1. \quad (109)$$

Let $\mathcal{P}_\Gamma(\{W^n, P_{S^n}\}_{n=1}^\infty)$ be the family of joint distributions $P_{U^n S^n X^n Y^n}$ such that for every $n \geq 1$, $U^n - (X^n, S^n) - Y^n$ forms a Markov chain, the \mathcal{S}^n -marginal of $P_{U^n S^n X^n Y^n}$ is P_{S^n} , the channel law $P_{Y^n | X^n, S^n} = W^n$ and

$$\mathfrak{p}\text{-}\limsup_{n \rightarrow \infty} g_n(X^n) \leq \Gamma \quad (110)$$

Define the quantity

$$\hat{C}_{\text{GP}}^*(\Gamma) := \sup \{ \underline{I}(\mathbf{U}; \mathbf{Y}) - \bar{I}(\mathbf{U}; \mathbf{S}) \} \quad (111)$$

where the supremum is over all joint distributions $\{P_{U^n S^n X^n Y^n}\}_{n=1}^\infty \in \mathcal{P}_\Gamma(\{W^n, P_{S^n}\}_{n=1}^\infty)$.

Theorem 23 (Lower Bound to the GP capacity [10]). *We have*

$$C_{\text{GP}}(\Gamma) \geq \hat{C}_{\text{GP}}^*(\Gamma). \quad (112)$$

Tan [10] also showed that the inequality in (112) is, in fact, tight. When the channel and state are discrete, stationary and memoryless, Tan [10] showed that the general formula in (111) reduces to the conventional one derived by Gel'fand-Pinsker [5] in (63). The proof of Theorem 23 parallels that for Theorem 22 and thus, we only sketch the proof by providing the choices of $|\mathcal{M}|$, L , γ_p and γ_c .

Proof: Fix an $\eta > 0$ and a sequence of joint distributions $\{P_{U^n S^n X^n Y^n}\}_{n=1}^\infty \in \mathcal{P}_\Gamma(\{W^n, P_{S^n}\}_{n=1}^\infty)$. Then make the following choices

$$\frac{1}{n} \log |\mathcal{M}| := \underline{I}(\mathbf{U}; \mathbf{Y}) - \bar{I}(\mathbf{U}; \mathbf{S}) - 4\eta \quad (113)$$

$$\frac{1}{n} \log L := \bar{I}(\mathbf{U}; \mathbf{S}) + 2\eta \quad (114)$$

$$\gamma_p := n(\underline{I}(\mathbf{U}; \mathbf{Y}) - \eta) \quad (115)$$

$$\gamma_c := n(\bar{I}(\mathbf{U}; \mathbf{S}) + \eta). \quad (116)$$

Because (110) holds for the sequence of joint distributions $\{P_{U^n S^n X^n Y^n}\}_{n=1}^\infty$, the average probability of error $P_e(\Phi_n; \Gamma + \eta)$ in (87) tends to zero and the rate of the code is given by (113). The proof is completed by again appealing to the *diagonal line argument* [7, Thm. 1.8.2]. ■

VI. ACHIEVABLE SECOND-ORDER CODING RATES

In this section, we demonstrate achievable second-order coding rates [11], [12], [24], [41], [42] for the three side-information problems of interest. Essentially, we are interested in characterizing the (n, ε) -optimal rate region for the WAK problem, the (n, ε) -Wyner-Ziv rate-distortion function and the (n, ε) -capacity of GP problem up to the second-order term. We do this by applying the multidimensional Berry-Esséen theorem [23], [52] to the finite blocklength CS-type bounds in Corollaries 15, 18 and 20. Throughout, we will not concern ourselves with optimizing the third-order terms.

The following important definition will be used throughout this section.

Definition 9. Let k be a positive integer. Let $\mathbf{V} \in \mathbb{R}^{k \times k}$ be a positive-semidefinite matrix that is not the all-zeros matrix but is allowed to be rank-deficient. Let the Gaussian random vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$. Define the set

$$\mathcal{S}(\mathbf{V}, \varepsilon) := \{\mathbf{z} \in \mathbb{R}^k : \Pr(\mathbf{Z} \leq \mathbf{z}) \geq 1 - \varepsilon\}. \quad (117)$$

This set was introduced in [25] and is, roughly speaking, the multidimensional analogue of the Q^{-1} function. Indeed, for $k = 1$ and any standard deviation $\sigma > 0$,

$$\mathcal{S}(\sigma^2, \varepsilon) = [\sigma Q^{-1}(\varepsilon), \infty). \quad (118)$$

Also, $\mathbf{1}_k$ and $\mathbf{0}_{k \times k}$ denote the length- k all-ones column vector and the $k \times k$ all-zeros matrix respectively.

A. Achievable Second-Order Coding Rates for the WAK problem

In this section, we derive an inner bound to $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ in (11) by the use of Gaussian approximations. Instead of simply applying the Berry-Esséen theorem to the information spectrum term within the simplified CS-type bound in (80), we enlarge our inner bound by using a “time-sharing” variable T , which is independent of (X, Y) . This technique was also used for the multiple access channel (MAC) by Huang and Moulin [45]. Note that in the finite blocklength setting, the region $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ does not have to be convex unlike in the asymptotic case; cf. (37). For fixed finite sets \mathcal{U} and \mathcal{T} , let $\tilde{\mathcal{P}}(P_{XY})$ be the set of all $P_{UTXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{T} \times \mathcal{X} \times \mathcal{Y})$ such that the $\mathcal{X} \times \mathcal{Y}$ -marginal of P_{UTXY} is P_{XY} , $U - (Y, T) - X$ forms a Markov chain and T is independent of (X, Y) .

Definition 10. The entropy-information density vector for the WAK problem for $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$ is defined as

$$\mathbf{j}(U, X, Y|T) := \begin{bmatrix} \log \frac{1}{P_{X|UT}(X|U, T)} \\ \log \frac{P_{Y|UT}(Y|U, T)}{P_Y(Y)} \end{bmatrix}. \quad (119)$$

Note that the mean of the entropy-information density vector in (119) is the vector of the entropy and mutual information, i.e.,

$$\mathbb{E}[\mathbf{j}(U, X, Y|T)] = \mathbf{J}(P_{UTXY}) = \begin{bmatrix} H(X|U, T) \\ I(U; Y|T) \end{bmatrix}. \quad (120)$$

The mutual information $I(U; Y|T) = I(U, T; Y)$ because T and Y are independent.

Definition 11. The entropy-information dispersion matrix for the WAK problem for a fixed $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$ is defined as

$$\mathbf{V}(P_{UTXY}) := \mathbb{E}_T [\text{Cov}(\mathbf{j}(U, X, Y|T))] \quad (121)$$

$$= \sum_{t \in \mathcal{T}} P_T(t) \text{Cov}(\mathbf{j}(U, X, Y|t)). \quad (122)$$

We abbreviate the deterministic quantities $\mathbf{J}(P_{UTXY}) \in \mathbb{R}_+^2$ and $\mathbf{V}(P_{UTXY}) \succeq 0$ as \mathbf{J} and \mathbf{V} respectively when the distribution $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$ is obvious from the context.

Definition 12. If $\mathbf{V}(P_{UTXY}) \neq \mathbf{0}_{2 \times 2}$, define $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY})$ to be the set of rate pairs (R_1, R_2) such that $\mathbf{R} := [R_1, R_2]^T$ satisfies

$$\mathbf{R} \in \mathbf{J} + \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_2. \quad (123)$$

If $\mathbf{V}(P_{UTXY}) = \mathbf{0}_{2 \times 2}$, define $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY})$ to be the set of rate pairs (R_1, R_2) such that

$$\mathbf{R} \in \mathbf{J} + \frac{2 \log n}{n} \mathbf{1}_2. \quad (124)$$

From the simplified CS-type bound for the WAK problem in Corollary 15, we can derive the following:

Theorem 24 (Inner Bound to (n, ε) -Optimal Rate Region). For every $0 < \varepsilon < 1$ and all n sufficiently large, the (n, ε) -optimal rate region $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ satisfies

$$\bigcup_{P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon). \quad (125)$$

From the modified CS-type bound for the WAK problem in Theorem 16, we can derive the following:

Theorem 25 (Modified Inner Bound to (n, ε) -Optimal Rate Region). *For every $0 < \varepsilon < 1$ and all n sufficiently large, the (n, ε) -optimal rate region $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ satisfies*

$$\bigcup_{P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})} \mathcal{R}'_{\text{in}}(n, \varepsilon; P_{UTXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon), \quad (126)$$

where $\mathcal{R}'_{\text{in}}(n, \varepsilon; P_{UTXY})$ is the set defined by replacing (123) with

$$\mathbf{R} \in \bigcup_{\rho \geq 0} \left\{ \mathbf{J} + \frac{\mathcal{S}(\mathbf{V}, \varepsilon) + [\rho, -\rho]^T}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_2 \right\}. \quad (127)$$

Remark 2. The bound in Theorem 25 is at least as tight as that in Theorem 24, and the former is strictly tighter than the latter for a fixed test channel. However, it is not clear whether the improvement is strict or not when we take the union over the test channels.

By setting $T = Y = U = \emptyset$ and $R_2 = 0$ in Theorem 25,⁴ we obtain a result first discovered by Strassen [41].

Corollary 26 (Achievable Second-Order Coding Rate for Lossless Source Coding). *Define the second-order coding rate for lossless source coding to be*

$$\sigma(P_X, \varepsilon) := \limsup_{n \rightarrow \infty} \sqrt{n}(R_X(n, \varepsilon) - H(X)) \quad (128)$$

where $R_X(n, \varepsilon)$ is the minimal rate of almost-lossless compression of source P_X at blocklength n with error probability not exceeding ε . Then,

$$\sigma(P_X, \varepsilon) \leq \sqrt{\text{Var}(\log P_X(X))} Q^{-1}(\varepsilon). \quad (129)$$

It is well-known that the result in Corollary 26 is tight, i.e., $\sqrt{\text{Var}(\log P_X(X))} Q^{-1}(\varepsilon)$ is the second-order coding rate for lossless source coding [11], [41], [42].

We refer to the reader to Appendix I for the proof of Theorem 24 (Appendix J for the proof of Theorem 25). The proof is based on the CS-type bound in (80) and the non-i.i.d. version of the multidimensional Berry-Esséen theorem by Götze [23]. The interpretation of this result is clear: From (123) which is the non-degenerate case, we see that the second-order coding rate region for a fixed P_{UTXY} is represented by the set $\mathcal{S}(\mathbf{V}(P_{UTXY}), \varepsilon)/\sqrt{n}$. Thus, the (n, ε) -optimal rate region converges to the asymptotic WAK region at a rate of $O(1/\sqrt{n})$ which can be predicted by the central limit theorem. More importantly, because our finite blocklength bound in (80) treats both the covering and binning error events *jointly*, this results in the coupling of the second-order rates through the set $\mathcal{S}(\mathbf{V}(P_{UTXY}), \varepsilon)$ and hence, the dispersion matrix $\mathbf{V}(P_{UTXY})$. This shows that the correlation between the entropy and information densities matters in the determination of the second-order coding rate.

More specifically, Theorems 24 and 25 are proved by taking $P_{U^n|Y^n}(u^n|y^n)$ to be equal to $P_{U|TY}^n(u^n|t^n, y^n)$ for some fixed (time-sharing) sequence $t^n \in \mathcal{T}^n$ and some joint distribution $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$. If $\mathcal{T} = \emptyset$, this is essentially using i.i.d. codes. An alternative to this proof strategy is to use conditionally constant composition codes as was done in Kelly-Wagner [33] to prove the error exponent result in Theorem 5. The advantage of this strategy is that it may yield better dispersion matrices because the unconditional dispersion matrix always dominates the conditional dispersion matrix [24, Lemma 62] (in the partial order induced by semi-definiteness). For using conditionally constant composition codes, we fix a conditional type $V_{Q_Y} \in \mathcal{V}_n(\mathcal{U}; Q_Y)$ for every marginal type $Q_Y \in \mathcal{P}_n(\mathcal{Y})$. Then, codewords are generated uniformly at random from $\mathcal{T}_{V_{Q_Y}}(y^n)$ if $y^n \in \mathcal{T}_{Q_Y}$. However, it does not appear that this strategy yields improved second-order coding rates compared to using i.i.d. codes as given in Theorems 24 and 25.

⁴In fact, to be precise, we cannot derive Corollary 26 from Theorem 24 because there is the residual term $\frac{2 \log n}{n}$ and we cannot set $R_2 = 0$. However, we can use Corollary 15 with $U = \emptyset$ to obtain Corollary 26 easily.

For comparison, for a fixed $P_{UXY} \in \mathcal{P}(P_{XY})$, define $\mathcal{R}_{\text{in}}^V(n, \varepsilon; P_{UXY})$ to be the set of rate pairs that satisfy

$$R_1 \geq H(X|U) + \sqrt{\frac{V_H(X|U)}{n}} Q^{-1}(\lambda\varepsilon) + \frac{2 \log n}{n} \quad (130)$$

$$R_2 \geq I(U; Y) + \sqrt{\frac{V_I(U; Y)}{n}} Q^{-1}((1-\lambda)\varepsilon) + \frac{2 \log n}{n} \quad (131)$$

for some $\lambda \in [0, 1]$ where the marginal entropy and information dispersions are defined as

$$V_H(X|U) := \text{Var} \left(\frac{1}{\log P_{X|U}(X|U)} \right) \quad (132)$$

$$V_I(U; Y) := \text{Var} \left(\log \frac{P_{Y|U}(Y|U)}{P_Y(Y)} \right) \quad (133)$$

respectively. Note that if $T = \emptyset$, then $V_H(X|U)$ and $V_I(U; Y)$ are the diagonal elements of the matrix $\mathbf{V}(P_{UTXY})$ in (121). It can easily be seen that Verdú's bound on the error probability of the WAK problem (42) yields the following inner bound on $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$.

$$\bigcup_{P_{UXY} \in \mathcal{P}(P_{XY})} \mathcal{R}_{\text{in}}^V(n, \varepsilon; P_{UXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon). \quad (134)$$

This “splitting” technique of ε into $\lambda\varepsilon$ and $(1-\lambda)\varepsilon$ in (130) and (131) was used by MolavianJazi and Laneman [46] in their work on finite blocklength analysis for the MAC. In Section VII, we numerically compare the inner bounds for the WAK problem provided in (125), (126) and (134).

Remark 3. From the non-asymptotic bound in Remark 1, we can also show that

$$\hat{\mathcal{R}}_{\text{in}}(n, \varepsilon) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon), \quad (135)$$

where $\hat{\mathcal{R}}_{\text{in}}(n, \varepsilon)$ is the set of rate pairs (R_1, R_2) such that

$$\begin{bmatrix} R_1 \\ R_1 + R_2 \end{bmatrix} \in \begin{bmatrix} H(X|Y) \\ H(X, Y) \end{bmatrix} + \frac{\mathcal{S}(\hat{\mathbf{V}}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_2 \quad (136)$$

for the covariance matrix

$$\hat{\mathbf{V}} = \text{Cov} \left(\begin{bmatrix} -\log P_{X|Y}(X|Y) \\ -\log P_{XY}(X, Y) \end{bmatrix} \right). \quad (137)$$

B. Achievable Second-Order Coding Rates for the WZ problem

In this section, we leverage on the simplified CS-type bound in Corollary 18 to derive an achievable second-order coding rate for the WZ problem. We do so by first finding an inner bound to the (n, ε) -Wyner-Ziv rate-distortion region $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$ defined in (17). Subsequently we find an upper bound to the (n, ε) -Wyner-Ziv rate-distortion function $R_{\text{WZ}}(n, \varepsilon)$ defined in (20). We also show that the (direct part of the) dispersion of lossy source coding found by Ingber-Kochman [27] and Kostina-Verdú [28] can be recovered from the CS-type bound in Corollary 18. This is not unexpected because the lossy source coding (rate-distortion) problem is a special case of the Wyner-Ziv problem where the side-information is absent.

We will again employ the “time-sharing” strategy used in Section VI-A. Note again that in the finite-blocklength setting $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$ does not have to be convex, unlike in the asymptotic setting. For fixed finite sets \mathcal{U} and \mathcal{T} , let $\mathcal{P}(P_{XY})$ be the collection of all pairs of joint distributions $P_{UTXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{T} \times \mathcal{X} \times \mathcal{Y})$ and functions $g : \mathcal{U} \times \mathcal{Y} \rightarrow \mathcal{X}$ such that the $\mathcal{X} \times \mathcal{Y}$ -marginal of P_{UTXY} is P_{XY} , $U - (X, T) - Y$ forms a Markov chain and T is independent of (X, Y) . Note that g does not necessarily have to satisfy the distortion constraint in (47).

In addition, to facilitate the time-sharing for distortion function, we define

$$d(X, g(U, Y)|T) := d(X_T, g(U_T, Y_T)) \quad (138)$$

where the random variables (U_t, X_t, Y_t) for any $t \in \mathcal{T}$ have distribution $P_{UXY|T=t}$.

Definition 13. The information-density-distortion vector for the WZ problem for $(P_{UTXY}, g) \in \tilde{\mathcal{P}}(P_{XY})$ is defined as

$$\mathbf{j}(U, X, Y|T) := \begin{bmatrix} -\log \frac{P_{Y|UT}(Y|U,T)}{P_Y(Y)} \\ \log \frac{P_{X|UT}(X|U,T)}{P_X(X)} \\ d(X, g(U, Y)|T) \end{bmatrix} \quad (139)$$

Since $\sum_t P_T(t) \mathbb{E}_{P_{U_{XY}|T}}[d(X_T, g(U_T, Y_T))|T = t] = \mathbb{E}[d(X, g(U, Y))]$, the expectation of information-density-distortion vector is given by

$$\mathbb{E}[\mathbf{j}(U, X, Y|T)] = \mathbf{J}(P_{UTXY}, g) = \begin{bmatrix} -I(U; Y|T) \\ I(U; X|T) \\ \mathbb{E}[d(X, g(U, Y))] \end{bmatrix} \quad (140)$$

Observe that the *sum* of the first two components of (140) resembles the Wyner-Ziv rate-distortion function defined in (48). As such when stating an achievable (n, ε) -Wyner-Ziv rate-distortion region, we project the first two terms onto an affine subspace representing their sum. See (143) and (144) below.

Definition 14. The information-distortion dispersion matrix for the WZ problem for a fixed $(P_{UTXY}, g) \in \tilde{\mathcal{P}}(P_{XY})$ is defined as

$$\mathbf{V}(P_{UTXY}, g) := \mathbb{E}_T [\text{Cov}(\mathbf{j}(U, X, Y|T))]. \quad (141)$$

Definition 15. Let $\mathbf{M} \in \mathbb{R}^{2 \times 3}$ be the matrix

$$\mathbf{M} := \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (142)$$

If $\mathbf{V}(P_{UTXY}, g) \neq \mathbf{0}_{3 \times 3}$, define $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, g)$ to be the set of all rate-distortion pairs (R, D) satisfying

$$\begin{bmatrix} R \\ D \end{bmatrix} \in \mathbf{M} \left(\mathbf{J} + \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_3 \right). \quad (143)$$

where $\mathbf{J} := \mathbf{J}(P_{UTXY}, g)$ and $\mathbf{V} := \mathbf{V}(P_{UTXY}, g)$. Else if $\mathbf{V}(P_{UTXY}, g) = \mathbf{0}_{3 \times 3}$, define $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, g)$ to be the set of all rate-distortion pairs (R, D) satisfying

$$\begin{bmatrix} R \\ D \end{bmatrix} \in \mathbf{M} \left(\mathbf{J} + \frac{2 \log n}{n} \mathbf{1}_3 \right). \quad (144)$$

In (143), the matrix \mathbf{M} serves project the three-dimensional set $\mathbf{J} + \mathcal{S}(\mathbf{V}, \varepsilon)/\sqrt{n} \subset \mathbb{R}^3$ onto two dimensions by linearly combining the first two mutual information terms to give $I(U; X|T) - I(U; Y|T) = I(U; X|Y, T)$ (by the Markov chain $U - (X, T) - Y$). From the simplified CS-type bound for the WZ problem in Corollary 18 and the multidimensional Berry-Esséen theorem [23], we can derive the following:

Theorem 27 (Inner Bound to the (n, ε) -Wyner-Ziv Rate-Distortion Region). *For every $0 < \varepsilon < 1$ and all n sufficiently large, the (n, ε) -Wyner-Ziv rate-distortion region $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$ satisfies*

$$\bigcup_{(P_{UTXY}, g) \in \tilde{\mathcal{P}}(P_{XY})} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, g) \subset \mathcal{R}_{\text{WZ}}(n, \varepsilon). \quad (145)$$

The proof of this result is provided in Appendix K. Further projecting onto the first dimension (the rate) for a fixed distortion level D yields the following:

Theorem 28 (Upper Bound to the (n, ε) -Wyner-Ziv Rate-Distortion Function). *For every $0 < \varepsilon < 1$ and all n sufficiently large, the (n, ε) -Wyner-Ziv rate-distortion function $R_{\text{WZ}}(n, \varepsilon, D)$ satisfies*

$$R_{\text{WZ}}(n, \varepsilon, D) \leq \inf \left\{ R : (R, D) \in \bigcup_{(P_{UTXY}, g) \in \tilde{\mathcal{P}}(P_{XY})} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, g) \right\}. \quad (146)$$

Theorems 27 and 28 are very similar in spirit to the result on the achievable second-order coding rate for the WAK problem. The marginal contributions from the distortion error event, the packing error event, the covering error event as well as their correlations are all involved in the dispersion matrix $\mathbf{V}(P_{UTXY}, g)$.

It is natural to wonder whether we are able to recover the dispersion for lossy source coding [27], [28] as a special case of Theorem 28 (like Corollary 26 is a special case of Theorem 25). This does not seem straightforward because of the distortion error event in (85). However, we can start from the CS-type bound in (85), set $Y = \emptyset$, $U = \hat{X}$ and use the method of types [30] or the notion of the D -tilted information [28] to obtain the specialization for the direct part. Before stating the result, we define a few quantities. Let the rate-distortion function of the source $X \sim Q \in \mathcal{P}(\mathcal{X})$ be denoted as

$$R(Q, D) := \min_{P_{\hat{X}, X}: P_X = Q, \mathbb{E}d(X, \hat{X}) \leq D} I(X; \hat{X}), \quad (147)$$

where $\mathbb{E}d(X, \hat{X}) := \sum_{x, \hat{x}} P_{\hat{X}, X}(\hat{x}, x) d(x, \hat{x})$. Also, define the D -tilted information to be

$$j(x, D) := -\log \mathbb{E} \left[\exp \left(\lambda^* D - \lambda^* d(x, \hat{X}^*) \right) \right] \quad (148)$$

where the expectation is with respect to the unconditional distribution of \hat{X}^* , the output distribution that optimizes the rate-distortion function in (147) and

$$\lambda^* := -\frac{\partial}{\partial D} R(P_X, D). \quad (149)$$

Theorem 29 (Achievable Second-Order Coding Rate for Lossy Source Coding). *Define the second-order coding rate for lossy source coding to be*

$$\sigma(P_X, D, \varepsilon) := \limsup_{n \rightarrow \infty} \sqrt{n} (R_X(n, \varepsilon; D) - R(P_X, D)) \quad (150)$$

where $R_X(n, \varepsilon; D)$ is the minimal rate of compression of source $X \sim P_X$ up to distortion D at blocklength n and probability of excess distortion not exceeding ε . We have

$$\sigma(P_X, D, \varepsilon) \leq \sqrt{\text{Var}(j(X, D))} Q^{-1}(\varepsilon) \quad (151)$$

Two proofs of Theorem 29 are provided in Appendix L, one based on the method of types and the other based on the D -tilted information in (148). For the former proof based on the method of types, we need to assume that $Q \mapsto R(Q, D)$ is differentiable in a small neighborhood of P_X and P_X is supported on a finite set. For the second proof, \mathcal{X} can be an abstract alphabet. Note that $R(P_X, D) = \mathbb{E}_{X \sim P_X} [j(X, D)]$. We remark that for discrete memoryless sources, the D -tilted information $j(x, D)$ coincides with the derivative of the rate-distortion function with respect to the source [27]

$$R'(x, D) = \frac{\partial}{\partial Q(x)} R(Q, D) \Big|_{Q=P_X}. \quad (152)$$

C. Achievable Second-Order Coding Rates for the GP problem

We conclude this section by stating an achievable second-order coding rate for the GP problem by presenting a lower bound to the (n, ε, Γ) -capacity $C_{\text{GP}}(n, \varepsilon, \Gamma)$ defined in (34). As in the previous two subsections, we start with definitions. For two finite sets \mathcal{U} and \mathcal{T} , define $\tilde{\mathcal{P}}(W, P_S)$ to be the collection of all $P_{TUSXY} \in \mathcal{P}(\mathcal{T} \times \mathcal{U} \times \mathcal{S} \times \mathcal{X} \times \mathcal{Y})$ such that the \mathcal{S} -marginal of P_{TUSXY} is P_S , $P_{Y|XS} = W$, $U - (X, S, T) - Y$ forms a Markov chain and T is independent of (S, X, Y) . Note that P_{TUSXY} does not necessarily have to satisfy the cost constraint in (62).

In addition, to facilitate the time-sharing for the cost function, we define

$$g(X|T) := g(X_T) \quad (153)$$

where X_t for any $t \in \mathcal{T}$ has distribution $P_{X|T=t}$.

Definition 16. The information-density-cost vector for the GP problem for $P_{TUSXY} \in \tilde{\mathcal{P}}(W, P_S)$ is defined as

$$\mathbf{j}(U, S, X, Y|T) := \begin{bmatrix} \log \frac{P_{Y|UT}(Y|U, T)}{P_Y(Y)} \\ -\log \frac{P_{S|UT}(S|U, T)}{P_S(S)} \\ -g(X|T) \end{bmatrix}. \quad (154)$$

Since $\sum_t P_T(t) \mathbb{E}_{P_{X|T}}[\mathbf{g}(X_T)|T=t] = \mathbb{E}[\mathbf{g}(X)]$, the expectation of this vector with respect to P_{TUSXY} is the vector of mutual informations and the negative cost, i.e.,

$$\mathbb{E}[\mathbf{j}(U, S, X, Y|T)] = \mathbf{J}(P_{TUSXY}) = \begin{bmatrix} I(U; Y|T) \\ -I(U; S|T) \\ -\mathbf{g}(X) \end{bmatrix}. \quad (155)$$

Definition 17. The information-dispersion matrix for the GP problem for $P_{TUSXY} \in \tilde{\mathcal{P}}(W, P_S)$ is defined as

$$\mathbf{V}(P_{TUSXY}) := \mathbb{E}_T[\text{Cov}(\mathbf{j}(U, S, X, Y|T))] \quad (156)$$

Definition 18. Let \mathbf{M} be the matrix defined in (142). If $\mathbf{V}(P_{TUSXY}) \neq \mathbf{0}_{3 \times 3}$, define the set $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{TUSXY})$ to be the set of all rate-cost pairs (R, Γ) satisfying

$$\begin{bmatrix} R \\ -\Gamma \end{bmatrix} \in \mathbf{M} \left(\mathbf{J} - \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} - \frac{2 \log n}{n} \mathbf{1}_3 \right) \quad (157)$$

where $\mathbf{J} := \mathbf{J}(P_{TUSXY})$ and $\mathbf{V} := \mathbf{V}(P_{TUSXY})$. Else if $\mathbf{V}(P_{TUSXY}) = \mathbf{0}_{3 \times 3}$, define $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{TUSXY})$ to be the set of all rate-cost pairs (R, Γ) satisfying

$$\begin{bmatrix} R \\ -\Gamma \end{bmatrix} \in \mathbf{M} \left(\mathbf{J} - \frac{2 \log n}{n} \mathbf{1}_3 \right). \quad (158)$$

By leveraging on our finite blocklength CS-type bound for the GP problem in (87), we obtain the following:

Theorem 30 (Inner Bound to the (n, ε) -GP Capacity-Cost Region). For every $0 < \varepsilon < 1$ and all n sufficiently large, the (n, ε) -GP capacity-cost region $\mathcal{C}_{\text{GP}}(n, \varepsilon)$ satisfies

$$\bigcup_{P_{TUSXY} \in \tilde{\mathcal{P}}(W, P_S)} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{TUSXY}) \subset \mathcal{C}_{\text{GP}}(n, \varepsilon). \quad (159)$$

By projecting onto the first dimension (the rate) for a fixed cost $\Gamma \geq 0$, we obtain:

Theorem 31 (Lower Bound to the (n, ε) -GP Capacity). For every $0 < \varepsilon < 1$ and all n sufficiently large, the (n, ε) -GP capacity-cost function $C_{\text{GP}}(n, \varepsilon, \Gamma)$ satisfies

$$C_{\text{GP}}(n, \varepsilon, \Gamma) \geq \sup \left\{ R : (R, \Gamma) \in \bigcup_{P_{TUSXY} \in \tilde{\mathcal{P}}(W, P_S)} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{TUSXY}) \right\} \quad (160)$$

The proof of Theorem 30 can be found in Appendix M. The matrix \mathbf{M} serves to project the first two components of each element in the set $\mathbf{J} + \mathcal{S}(\mathbf{V}, \varepsilon)/\sqrt{n}$ onto one dimension. Indeed, for a fixed $P_{TUSXY} \in \tilde{\mathcal{P}}(W, P_S)$, the first two components read $I(U; Y|T) - I(U; S|T)$ which, if $T = \emptyset$ and the random variables (U, S, X, Y) are capacity-achieving, reduces to the GP formula in (63). Hence, the set $\mathbf{M}\mathcal{S}(\mathbf{V}, \varepsilon)/\sqrt{n} \subset \mathbb{R}$ quantifies all possible backoffs from the asymptotic GP capacity-cost region \mathcal{C}_{GP} (defined in (33)) at blocklength n and average error probability ε based on our CS-type finite blocklength bound for the GP problem in (87). The bound in (160) is clearly much tighter than the one provided in [10] which is based on the use of Wyner's PBL and Markov lemma.

Now by setting $S = T = \emptyset$, $U = X$ and $\Gamma = \infty$ in Theorem 31, we recover the direct part of the second-order coding rate for channel coding without cost constraints [12], [24], [41].

Corollary 32 (Achievable Second-Order Coding Rate for Channel Coding). Fix a non-exotic [24] discrete memoryless channel $W : \mathcal{X} \rightarrow \mathcal{Y}$ with channel capacity $C(W) = \max_{P_X} I(X; Y)$. Define the second-order coding rate for channel coding to be

$$\sigma(W, \varepsilon) := \limsup_{n \rightarrow \infty} \sqrt{n}(C(W) - C_W(n, \varepsilon)) \quad (161)$$

where $C_W(n, \varepsilon)$ is the maximal rate of transmission over the channel W at blocklength n and average error probability ε . Then,

$$\sigma(W, \varepsilon) \leq \min_{P_{X^*}} \sqrt{\text{Var} \left(\log \frac{W(Y^*|X^*)}{P_{Y^*}(Y^*)} \right)} Q^{-1}(\varepsilon) \quad (162)$$

where $(X^*, Y^*) \sim P_{X^*} \times W$ and the minimization is over all capacity-achieving input distributions.

The bound in (162) has long been known to be an equality [41]. Note that the unconditional dispersion in (162) $\text{Var} \left(\log \frac{W(Y^*|X^*)}{P_{Y^*}(Y^*)} \right)$ coincides with the conditional dispersion [24] since it is being evaluated at a capacity-achieving input distribution. As such, the converse can be proved using the meta-converse in [24] or an modification of the Verdú-Han converse [7, Lem. 3.2.2] with an judiciously chosen output distribution as was done in [12]. In fact, we can also derive a generalization of Corollary 32 with cost constraints incorporated [12, Thm. 3] using similar techniques as in the proof of Theorem 29. Namely, we use a uniform distribution over a particular type class (constant composition codes) as the input distribution. The type is chosen to be close to the optimal input distribution (assuming it is unique).

VII. NUMERICAL EXAMPLES

A. Numerical Example for WAK Problem

In this section, we use an example to illustrate the inner bound on (n, ε) -optimal rate region for the WAK problem obtained in Theorem 24. We neglect the small $O \left(\frac{\log n}{n} \right)$ term. The source is taken to be a discrete symmetric binary source DSBS(α), i.e.,

$$P_{XY} = \frac{1}{2} \begin{bmatrix} 1 - \alpha & \alpha \\ \alpha & 1 - \alpha \end{bmatrix}. \quad (163)$$

In this case, the optimal rate region reduces to

$$\mathcal{R}_{\text{WAK}}^* = \left\{ (R_1, R_2) : R_1 \geq h(\beta * \alpha), R_2 \geq 1 - h(\beta), 0 \leq \beta \leq \frac{1}{2} \right\}, \quad (164)$$

where $h(\cdot)$ is the binary entropy function and $\beta * \alpha := \beta(1 - \alpha) + (1 - \beta)\alpha$ is the binary convolution. The above region is attained by setting the backward test channel from U to Y to be a BSC with some crossover probability β . All the elements in the entropy-information dispersion matrix $\mathbf{V}(\beta)$ can be evaluated in closed form in terms of β . Define $\mathbf{J}(\beta) := [h(\beta * \alpha), 1 - h(\beta)]^T$. In Fig. 4, we plot the second-order region

$$\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon) := \bigcup_{0 \leq \beta \leq \frac{1}{2}} \left\{ (R_1, R_2) : \mathbf{R} \in \mathbf{J}(\beta) + \frac{\mathcal{S}(\mathbf{V}(\beta), \varepsilon)}{\sqrt{n}} \right\}. \quad (165)$$

The first-order region $\mathcal{R}_{\text{WAK}}^*$ and the second-order region with simple time-sharing ($|\mathcal{T}| = 2$) are also shown for comparison. More precisely, the simple time-sharing is between $\beta = 0$ and $\beta = 1/2$. As expected, as the block length increases, the (n, ε) -optimal rate region tends to the first-order one. Interestingly, at small block length, time-sharing makes the second-order (n, ε) -optimal rate region in (165) larger compared to that without time-sharing. Especially, the simple time-sharing is better than $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$ for $n = 500$ because the rank of the entropy-information dispersion matrix $\lambda \mathbf{V}(0) + (1 - \lambda) \mathbf{V}(1/2)$ for $0 < \lambda \leq 1$ is one.⁵

We also consider the region $\tilde{\mathcal{R}}_{\text{in}}^{\text{V}}(n, \varepsilon)$ which is the analogue of $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$ but derived from Verdú's bound in Theorem 4. In Fig. 5, we compare the second-order coefficients, namely that derived from our bound $\mathcal{S}(\mathbf{V}(\beta), \varepsilon)$ and

$$\mathcal{S}^{\text{V}}(\mathbf{V}(\beta), \varepsilon) := \bigcup_{0 \leq \lambda \leq 1} \left\{ (z_1, z_2) : z_1 \geq \sqrt{V_H(\beta)} Q^{-1}(\lambda \varepsilon), z_2 \geq \sqrt{V_I(\beta)} Q^{-1}((1 - \lambda) \varepsilon) \right\}. \quad (166)$$

Note that the difference between the two regions is quite small even for $\varepsilon = 0.5$. This is because, for this example, the covariance of the entropy- and information-density (off-diagonal in the dispersion matrix) is negative so the difference between $\Pr(Z_1 \geq z_1 \text{ or } Z_2 \geq z_2)$ and $\Pr(Z_1 \geq z_1) + \Pr(Z_2 \geq z_2)$ is small. In this case, the 2-dimensional Gaussian $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}(\beta))$ has a negative covariance and hence the probability mass in the first and third quadrants are small. Hence, the union bound is not very loose in this case.

Next, we consider the binary joint source given by $P_{X|Y}(1|0) = P_{X|Y}(0|1) = \alpha$ and $P_Y(0) = p \leq \frac{1}{2}$, which is a generalization of (163). This example was investigated in [53], and the optimal rate region reduces to

$$\mathcal{R}_{\text{WAK}}^* = \{(R_1, R_2) : R_1 \geq h(\beta * \alpha), R_2 \geq h(p) - h(\beta), 0 \leq \beta \leq p\}. \quad (167)$$

⁵It should be noted that the rank of $\mathbf{V}(1/2)$ is zero.

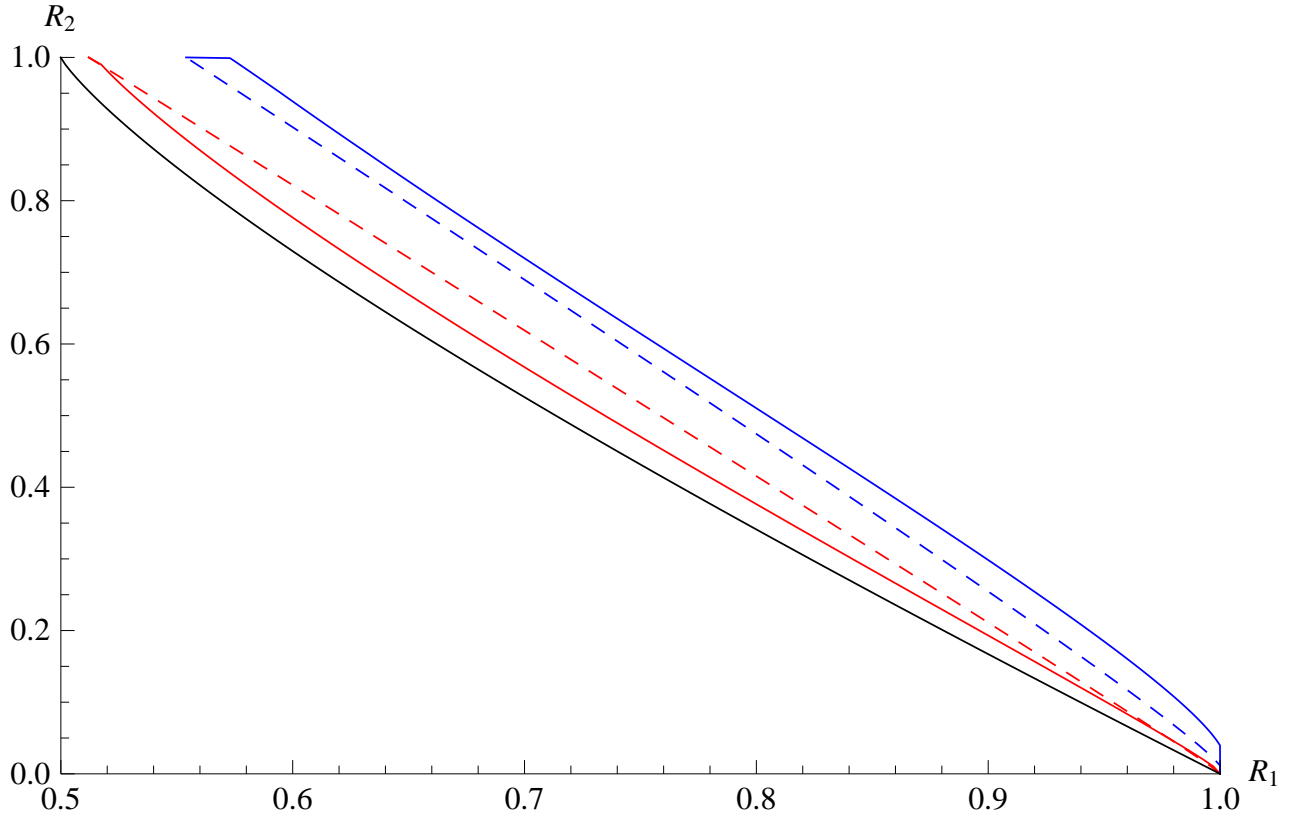


Fig. 4. A comparison between $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$ without time-sharing (solid line) and the time-sharing region (dashed line) for $\varepsilon = 0.1$. The regions are to the top right of the curves. The blue and red curves are for $n = 500$ and $n = 10,000$ respectively. The black curve is the first-order region (1).

The above region is attained by setting the backward test channel from U to Y to be BSC with some crossover probability $0 \leq \beta \leq p$. All the elements in the entropy-information dispersion matrix $\mathbf{V}(\beta)$ can be evaluated in closed form in terms of β . Define $\mathbf{J}(\beta) := [h(\beta * \alpha), h(p) - h(\beta)]^T$. In Fig. 6, we plot the second-order region

$$\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon) := \bigcup_{0 \leq \beta \leq p} \left\{ (R_1, R_2) : \mathbf{R} \in \mathbf{J}(\beta) + \frac{\mathcal{S}(\mathbf{V}(\beta), \varepsilon)}{\sqrt{n}} \right\}. \quad (168)$$

For comparison, we also plot the second-order region derived from Remark 3. Around the corner point defined by the entropies $[H(X|Y), H(Y)]^T = [h(\beta), h(p)]^T$, we find that the bound from Remark 3 is tighter than that given by (168).

B. Numerical Example for GP Problem

In this section, we use an example to illustrate the inner bound on (n, ε) -optimal rate for the GP problem obtained in Theorem 30. We do not consider cost constraints here, i.e., $\Gamma = \infty$. We also neglect the small $O\left(\frac{\log n}{n}\right)$ term. We consider the *memory with stuck-at faults* example [54] (see also [1, Example 7.3]). The state $S = 0$ correspond to a faculty memory cell that output 0 independent of the input value, the state $S = 1$ corresponds to a faculty memory cell that outputs 1 independent of the input value, and the state $S = 2$ corresponds to a binary symmetric channel with crossover probability α . The probabilities of these states are $\frac{p}{2}$, $\frac{p}{2}$, and $1 - p$ respectively.

It is known [54] that the capacity is

$$C_{\text{GP}}^* = (1 - p)(1 - h(\alpha)). \quad (169)$$

The above capacity is attained by setting $\mathcal{U} = \{0, 1\}$ and $P_{U|X}(0|0) = P_{U|S}(1|1) = 1 - \alpha$, $P_{U|S}(u|2) = \frac{1}{2}$, and $X = U$. All the elements in the information dispersion matrix \mathbf{V} can be evaluated in closed form. In Fig. 7, we

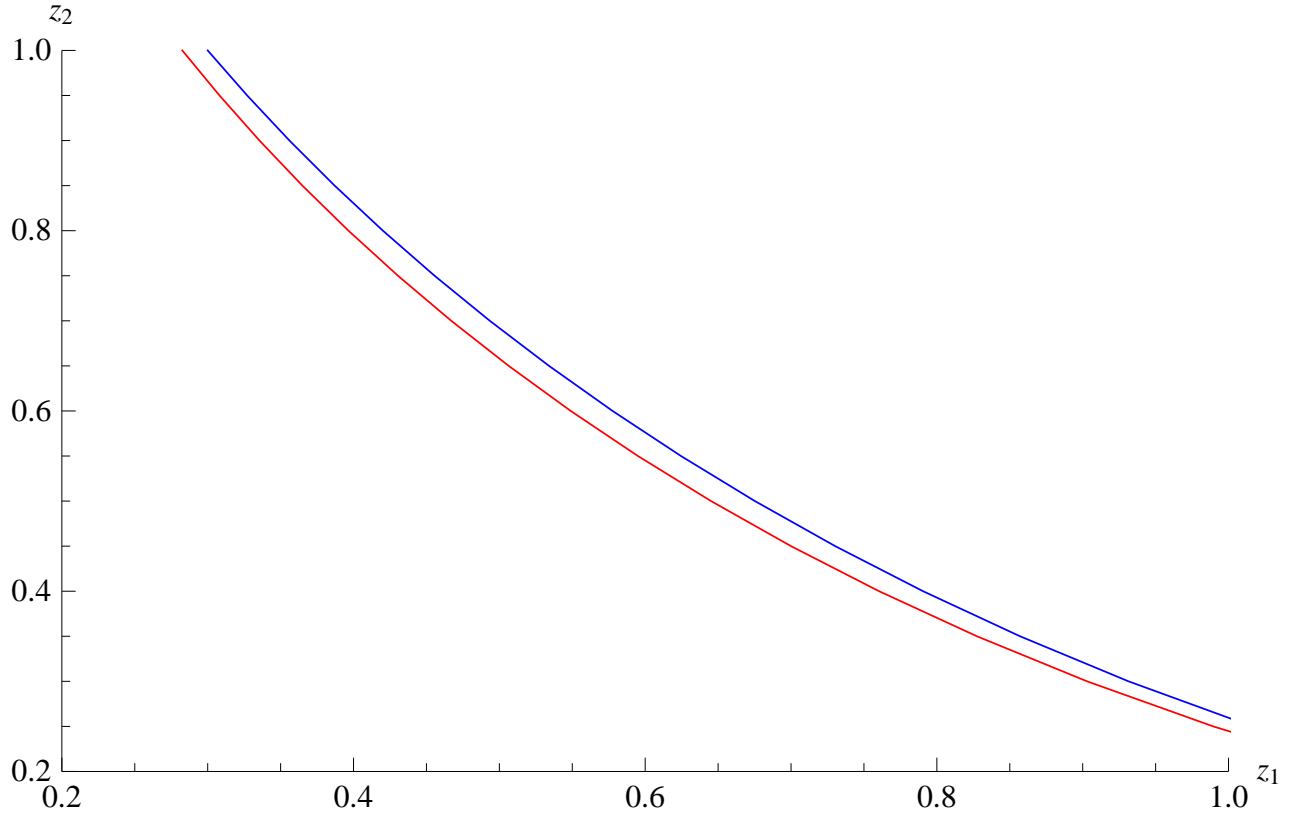


Fig. 5. A comparison between $\mathcal{S}(\mathbf{V}(\beta), \varepsilon)$ (defined in (117)) and $\mathcal{S}^V(\mathbf{V}(\beta), \varepsilon)$ (defined in (166)) for $\beta = h^{-1}(0.5)$ and $\varepsilon = 0.5$. The red and blue curves are the boundaries of $\mathcal{S}(\mathbf{V}(\beta), \varepsilon)$ and $\mathcal{S}^V(\mathbf{V}(\beta), \varepsilon)$ respectively. The regions lie to the top right of the curves.

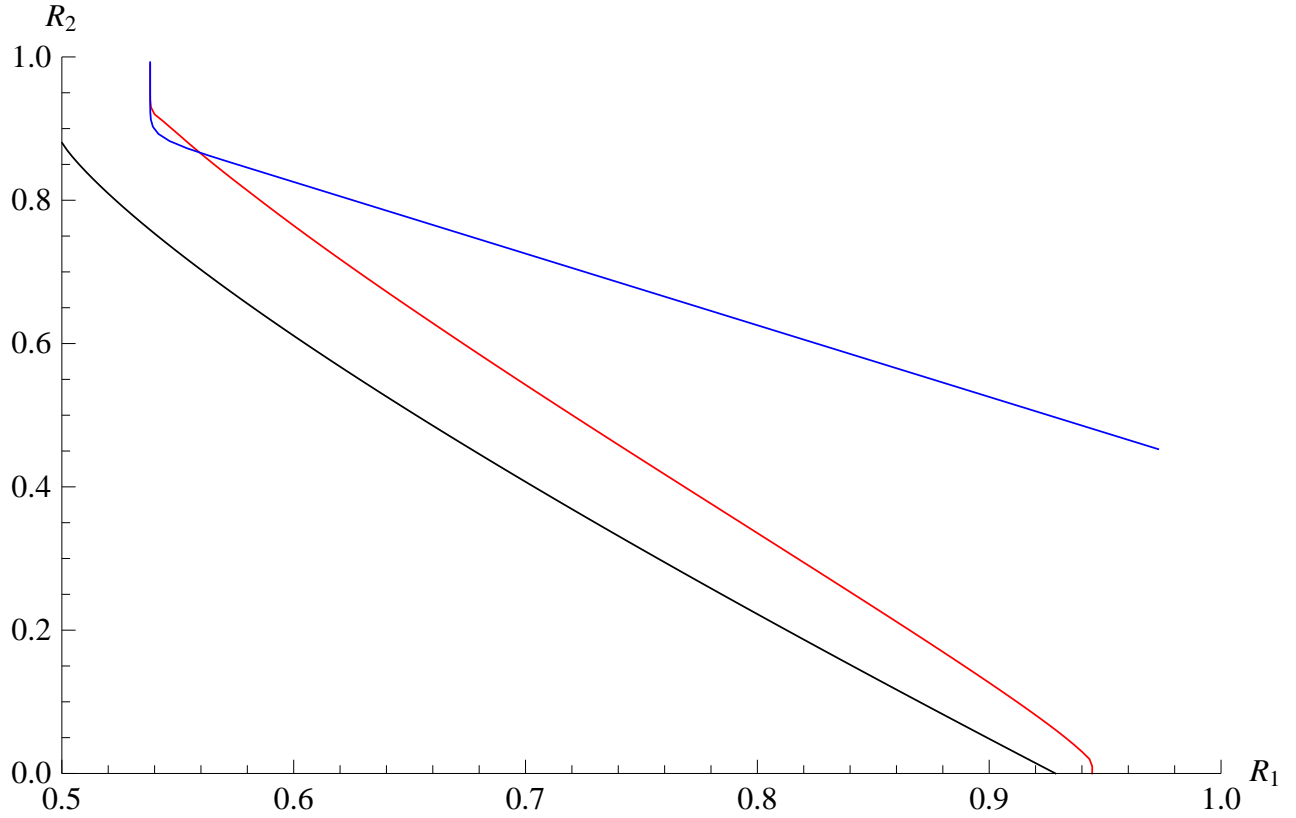


Fig. 6. A comparison between $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$ (red solid curve) and the bound from Remark 3 (blue solid curve) for $\varepsilon = 0.1$ and $n = 1000$. The regions are to the top right of the curves.

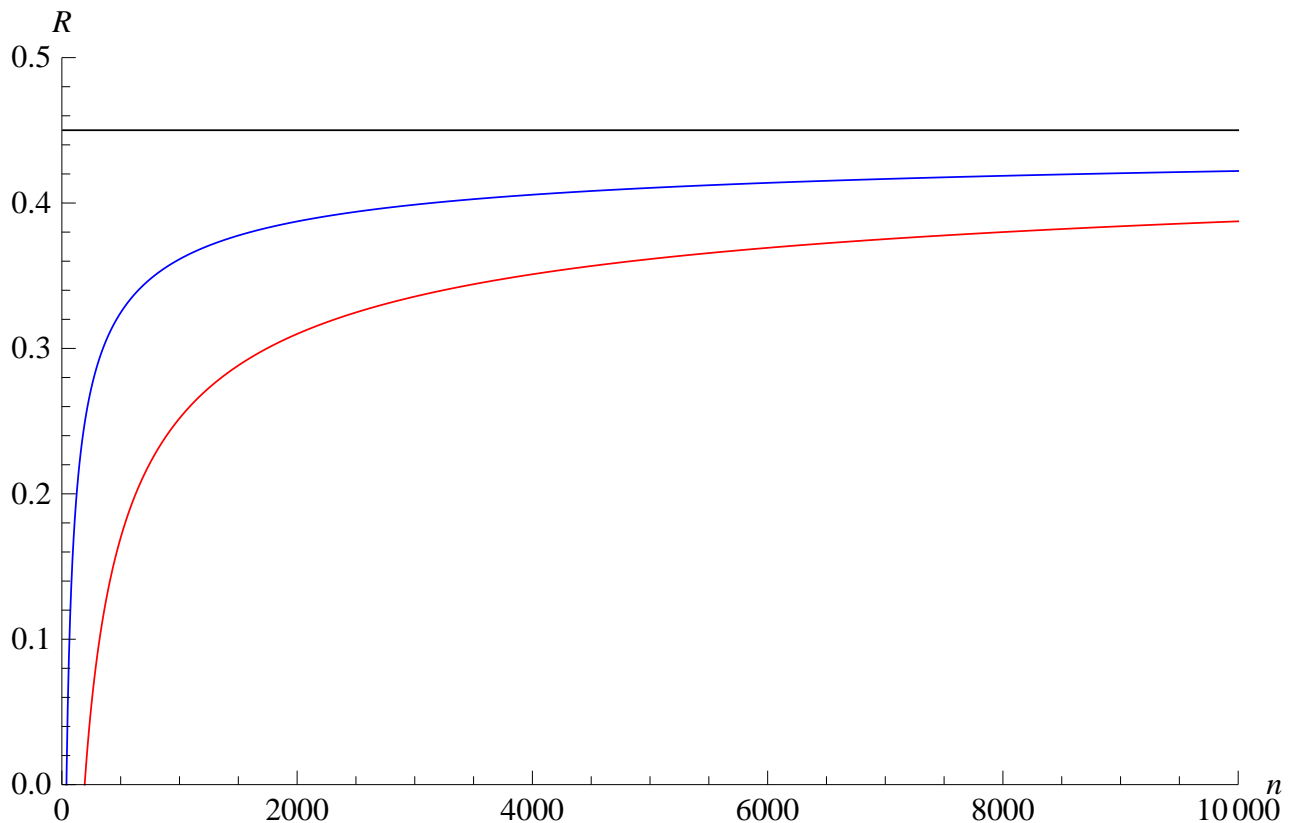


Fig. 7. A comparison between $\tilde{R}_{\text{GP}}(n, \varepsilon; p, \alpha)$ (red solid line) and $\tilde{C}(n, \varepsilon; p, \alpha)$ (blue solid line) for $\varepsilon = 0.001$, $p = 0.1$, and $\alpha = 0.11$. The black solid line is the first-order capacity (169).

plot the second-order capacity

$$\tilde{R}_{\text{GP}}(n, \varepsilon; p, \alpha) := (1 - p)(1 - h(\alpha)) - \frac{1}{\sqrt{n}} \min\{z_1 + z_2 : (z_1, z_2) \in \mathcal{S}(\mathbf{V}, \varepsilon)\}. \quad (170)$$

For comparison, let us consider the case in which the decoder, instead of the encoder, can access the state S . In this case, we can regard X as the channel input and (S, Y) as the channel output. It is known [54] that the capacity $C(W)$ of this channel is the same as (169). The dispersion V can be evaluated in closed form by appealing to the law of total variance [55]. In Fig. 7, we also plot the second order capacity

$$\tilde{C}(n, \varepsilon; p, \alpha) := (1 - p)(1 - h(\alpha)) - \sqrt{\frac{V}{n}} Q^{-1}(\varepsilon). \quad (171)$$

From the figure, we can find that the lower bound $\tilde{R}_{\text{GP}}(n, \varepsilon; p, \alpha)$ on the GP (n, ε) -optimal rate is smaller than the (n, ε) -optimal rate with decoder side-information though the first order rates coincide.

VIII. CONCLUSION AND FURTHER WORK

In this paper, we proved several new non-asymptotic bounds on the error probability for side-information coding problems, including the WAK, WZ and GP problems. These bounds then yield known general formulas as simple corollaries. In addition, we used these bounds to provide achievable second-order coding rates for these three side-information problems. We argued that when evaluated using i.i.d. test channels, the second-order rates resulting from our non-asymptotic bounds are the best known in the literature. In particular, they improve on Verdú's work on non-asymptotic achievability bounds for multi-user information theory problems [6]. Other challenging problems along the same lines include the Heegard-Berger [56] problem, multiple description coding [57], Marton's inner bound for the broadcast channel [58], [59], compress-forward for the relay channel [60] and hypothesis testing with multi-terminal data compression [61].

APPENDIX A
PROOF OF PROPOSITION 1 (EXPURGATED CODE)

Proof: Let $x_0 \in \mathcal{X}$ be a prescribed constant satisfying $g(x_0) \leq \Gamma$, and let P_X^* be the distribution such that $P_X^*(x_0) = 1$, i.e., $P_X^*(x) = \mathbf{1}[x = x_0]$. Then, we define

$$\tilde{P}_{X|MS}(x|m, s) := P_{X|MS}(x|m, s)\mathbf{1}[g(x) \leq \Gamma] + P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c|m, s)P_X^*(x). \quad (172)$$

Then, it is obvious that $\tilde{P}_X(\mathcal{T}_g^{\text{GP}}(\Gamma)) = 1$. We also have

$$\begin{aligned} & \tilde{P}_{MSXY\hat{M}}[m \neq \hat{m}] \\ &= \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m)P_S(s)\tilde{P}_{X|MS}(x|m, s)W(y|x, s)P_{\hat{M}|Y}(\hat{m}|y) \end{aligned} \quad (173)$$

$$\begin{aligned} &= \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m)P_S(s)P_{X|MS}(x|m, s)W(y|x, s)P_{\hat{M}|Y}(\hat{m}|y)\mathbf{1}[g(x) \leq \Gamma] \\ &\quad + \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m)P_S(s)P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c|m, s)P_X^*(x)W(y|x, s)P_{\hat{M}|Y}(\hat{m}|y) \end{aligned} \quad (174)$$

$$\begin{aligned} &\leq \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m)P_S(s)P_{X|MS}(x|m, s)W(y|x, s)P_{\hat{M}|Y}(\hat{m}|y)\mathbf{1}[g(x) \leq \Gamma] \\ &\quad + \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m)P_S(s)P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c|m, s)P_X^*(x)W(y|x, s)P_{\hat{M}|Y}(\hat{m}|y) \end{aligned} \quad (175)$$

$$\begin{aligned} &= \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m)P_S(s)P_{X|MS}(x|m, s)W(y|x, s)P_{\hat{M}|Y}(\hat{m}|y)\mathbf{1}[g(x) \leq \Gamma] \\ &\quad + \sum_{m, s} P_M(m)P_S(s)P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c|m, s) \end{aligned} \quad (176)$$

$$= P_{MSXY\hat{M}}[g(x) \leq \Gamma \cap m \neq \hat{m}] + P_{MSXY\hat{M}}[g(x) > \Gamma] \quad (177)$$

$$= P_{MSXY\hat{M}}[g(x) > \Gamma \cup m \neq \hat{m}] \quad (178)$$

as desired. ■

APPENDIX B
CHANNEL RESOLVABILITY

In this appendix, we review notations and known results for channel resolvability [7, Ch. 6] [13] [14].

As a start, we first review the properties of the variational distance. Let $\mathcal{P}'(\mathcal{U})$ be the set of all sub-normalized non-negative functions (not necessarily probability distribution unless otherwise stated) on a finite set \mathcal{U} . Note that if $P \in \mathcal{P}'(\mathcal{U})$ is normalized then $P \in \mathcal{P}(\mathcal{U})$, i.e., P is a distribution on \mathcal{U} . For $P, Q \in \mathcal{P}'(\mathcal{U})$, we define the variational distance (divided by 2) as

$$d(P, Q) = \frac{1}{2} \sum_{u \in \mathcal{U}} |P(u) - Q(u)|. \quad (179)$$

For two sets \mathcal{U} and \mathcal{Z} , let $\mathcal{P}'(\mathcal{Z}|\mathcal{U})$ be the set of all sub-normalized non-negative functions indexed by $u \in \mathcal{U}$. When $W \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$ is normalized, it is a channel. In this section, we denote the joint distribution induced by $P \in \mathcal{P}(\mathcal{U})$ and $W \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$ as $PW \in \mathcal{P}'(\mathcal{U} \times \mathcal{Z})$. The following properties are useful in the proof of theorems.

Lemma 33. *The variational distance satisfies the following properties.*

- 1) *The monotonicity with respect to marginalization: For $P, Q \in \mathcal{P}'(\mathcal{U})$ and $W, V \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$, let $P', Q' \in \mathcal{P}'(\mathcal{Z})$ be*

$$P'(z) := \sum_{u \in \mathcal{U}} P(u)W(z|u), \quad Q'(z) := \sum_{u \in \mathcal{U}} Q(u)V(z|u). \quad (180)$$

Then,

$$d(P', Q') \leq d(PW, QV). \quad (181)$$

2) The data-processing inequality: For $P, Q \in \mathcal{P}'(\mathcal{U})$ and $W \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$,

$$d(PW, QW) \leq d(P, Q). \quad (182)$$

3) For a distribution $P \in \mathcal{P}(\mathcal{U})$, a sub-normalized measure $Q \in \mathcal{P}'(\mathcal{U})$, and any subset $\Gamma \subset \mathcal{U}$,

$$P(\Gamma) \leq Q(\Gamma) + d(P, Q) + \frac{1 - Q(\mathcal{U})}{2}. \quad (183)$$

Remark 4. Combining (181) and (182), we have

$$d(P', Q') \leq d(P, Q). \quad (184)$$

Although the above inequality is usually referred as the data-processing inequality, we will use (182) in the proofs of non-asymptotic bounds.

Proof: Since

$$d(P', Q') = \frac{1}{2} \sum_z \left| \sum_u P(u)W(z|u) - \sum_u Q(u)V(z|u) \right| \quad (185)$$

$$\leq \frac{1}{2} \sum_z \sum_u |P(u)W(z|u) - Q(u)V(z|u)| \quad (186)$$

$$= d(PW, QV) \quad (187)$$

holds, we have (181). On the other hand, we have

$$d(PW, QW) = \frac{1}{2} \sum_{u,z} |P(u)W(z|u) - Q(u)W(z|u)| \quad (188)$$

$$= \frac{1}{2} \sum_{u,z} W(z|u) |P(u) - Q(u)| \quad (189)$$

$$\leq \frac{1}{2} \sum_u |P(u) - Q(u)| \quad (190)$$

$$= d(P, Q) \quad (191)$$

and thus, (182) holds.

Further, letting $\mathcal{U}_P = \{u : P(u) \geq Q(u)\}$ and $q = 1 - Q(\mathcal{U})$, we have

$$d(P, Q) = \frac{1}{2} \sum_u |P(u) - Q(u)| \quad (192)$$

$$= \frac{1}{2} [\{P(\mathcal{U}_P) - Q(\mathcal{U}_P)\} + \{Q(\mathcal{U}_P^c) - P(\mathcal{U}_P^c)\}] \quad (193)$$

$$= \frac{1}{2} [\{P(\mathcal{U}_P) - Q(\mathcal{U}_P)\} + \{1 - P(\mathcal{U}_P^c) - (Q(\mathcal{U}) - Q(\mathcal{U}_P^c)) - q\}] \quad (194)$$

$$= P(\mathcal{U}_P) - Q(\mathcal{U}_P) - \frac{q}{2} \quad (195)$$

$$\geq P(\Gamma) - Q(\Gamma) - \frac{q}{2}. \quad (196)$$

Hence, we have (183). ■

Next, we introduce the concept of *smoothing* of a distribution [62]. For a distribution $P \in \mathcal{P}(\mathcal{U})$ and a subset $\mathcal{T} \subset \mathcal{U}$, a smoothed sub-normalized function \bar{P} of P is derived by

$$\bar{P}(u) := P(u)\mathbf{1}[u \in \mathcal{T}]. \quad (197)$$

Note that the distance between the original distribution and a smoothed one is

$$d(P, \bar{P}) = \frac{P(\mathcal{T}^c)}{2}. \quad (198)$$

Similarly, for a channel $W: \mathcal{U} \rightarrow \mathcal{Z}$ and a subset $\mathcal{T} \subset \mathcal{U} \times \mathcal{Z}$, a smoothed one $\bar{W} \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$ is derived by

$$\bar{W}(z|u) := W(z|u)\mathbf{1}[(u, z) \in \mathcal{T}] \quad (199)$$

and it satisfies

$$d(PW, P\bar{W}) = \frac{PW(\mathcal{T}^c)}{2}, \quad (200)$$

where $PW \in \mathcal{P}(\mathcal{U} \times \mathcal{Z})$ is the joint distribution induced by P and W .

Now, we consider the problem of channel resolvability. Let a channel $P_{Z|U}: \mathcal{U} \rightarrow \mathcal{Z}$ and an input distribution P_U be given. We would like to approximate the output distribution

$$P_Z(z) = \sum_{u \in \mathcal{U}} P_U(u)P_{Z|U}(z|u) \quad (201)$$

by using $P_{Z|U}$ and as small an amount of randomness as possible. This is done by means of a designing a deterministic map from a finite set \mathcal{I} to a codebook $\mathcal{C} = \{u_i\}_{i \in \mathcal{I}} \subset \mathcal{U}$. For a given resolvability code \mathcal{C} , let

$$P_{\bar{Z}}(z) = \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}|} P_{Z|U}(z|u_i) \quad (202)$$

be the simulated output distribution. The approximation error is evaluated by the distance $d(P_{\bar{Z}}, P_Z)$.

We consider using the random coding technique as follows. We randomly and independently generate codewords $u_1, u_2, \dots, u_{|\mathcal{I}|}$ according to P_U . To derive an upper bound on the averaged approximation error $\mathbb{E}_{\mathcal{C}} [d(P_{\bar{Z}}, P_Z)]$, it is convenient to consider a smoothing operation defined as follows. For the set

$$\mathcal{T}_c(\gamma_c) := \left\{ (u, z) : \log \frac{P_{Z|U}(z|u)}{P_Z(z)} \leq \gamma_c \right\}, \quad (203)$$

let

$$\bar{P}_{Z|U}(z|u) := P_{Z|U}(z|u)\mathbf{1}[(u, z) \in \mathcal{T}_c(\gamma_c)]. \quad (204)$$

Moreover, for fixed resolvability code $\mathcal{C} = \{u_1, \dots, u_{|\mathcal{I}|}\}$, let

$$\bar{P}_{\bar{Z}}(z) := \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \bar{P}_{Z|U}(z|u_i). \quad (205)$$

Then, we have the following lemma.

Lemma 34 (Lemma 2 of [14]). *For any $\gamma_c \geq 0$, we have*

$$\mathbb{E}_{\mathcal{C}} [d(\bar{P}_{\bar{Z}}, \bar{P}_Z)] \leq \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UZ})}{|\mathcal{I}|}} \quad (206)$$

and

$$\mathbb{E}_{\mathcal{C}} [d(P_{\bar{Z}}, P_Z)] \leq P_{UZ}(\mathcal{T}_c(\gamma_c)^c) + \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UZ})}{|\mathcal{I}|}}, \quad (207)$$

where $\bar{P}_Z(z) = \sum_u P_U(u)P_{Z|U}(z|u)$.

Remark 5. Although the statement of [14, Lemma 2] is that (207) holds, (206) is also proved in the proof of [14, Lemma 2] (at left bottom of [14, pp. 1566]).

The definition of $l_X(y)$ in the proof of [14, Lemma 2] is incorrect. Hence, for completeness, we provide the proof of (206) in Lemma 34.

Proof: By Jensen's inequality and the convexity of $t \mapsto t^2$,

$$\mathbb{E}_{\mathcal{C}} [d(\bar{P}_{\bar{Z}}, \bar{P}_Z)]^2 \leq \mathbb{E}_{\mathcal{C}} [d(\bar{P}_{\bar{Z}}, \bar{P}_Z)^2] = \frac{1}{4} \mathbb{E}_{\mathcal{C}} [\|\bar{P}_{\bar{Z}} - \bar{P}_Z\|_1^2], \quad (208)$$

where $\|x\|_1 := \sum_i |x_i|$ is the ℓ_1 norm. We now bound the term on the right as follows:

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} [\|\bar{P}_{\bar{Z}} - \bar{P}_Z\|_1^2] \\ &= \mathbb{E}_{\mathcal{C}} \left[\left(\sum_z |\bar{P}_{\bar{Z}}(z) - \bar{P}_Z(z)| \right)^2 \right] \end{aligned} \quad (209)$$

$$= \mathbb{E}_{\mathcal{C}} \left[\left(\sum_z \sqrt{P_Z(z)} \sqrt{P_Z(z)} \left| \frac{\bar{P}_{\bar{Z}}(z) - \bar{P}_Z(z)}{P_Z(z)} \right| \right)^2 \right] \quad (210)$$

$$\leq \mathbb{E}_{\mathcal{C}} \left[\sum_z P_Z(z) \left| \frac{\bar{P}_{\bar{Z}}(z) - \bar{P}_Z(z)}{P_Z(z)} \right|^2 \right] \quad (211)$$

$$= \mathbb{E}_{\mathcal{C}, P_Z} \left[\left| \frac{\bar{P}_{\bar{Z}} - \bar{P}_Z}{P_Z} \right|^2 \right] \quad (212)$$

where (211) follows from Cauchy-Schwarz inequality regarding $a_z := \sqrt{P_Z(z)}$ as one $|\mathcal{Z}|$ -dimensional vector and $b_z := \sqrt{P_Z(z)} \left| \frac{\bar{P}_{\bar{Z}}(z) - \bar{P}_Z(z)}{P_Z(z)} \right|$ as another $|\mathcal{Z}|$ -dimensional vector. In fact, $|\mathcal{Z}|$ does not have to be finite for the Cauchy-Schwarz inequality to be valid. Continuing, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} [\|\bar{P}_{\bar{Z}} - \bar{P}_Z\|_1^2] \\ & \leq \mathbb{E}_{P_Z} \mathbb{E}_{\mathcal{C}} \left[\left(\frac{\bar{P}_{\bar{Z}}(z)}{P_Z(z)} - \frac{\bar{P}_Z(z)}{P_Z(z)} \right)^2 \right] \end{aligned} \quad (213)$$

$$= \mathbb{E}_{P_Z} \mathbb{E}_{\mathcal{C}} \left[\left(\sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \frac{\bar{P}_{Z|U}(z|u_i)}{P_Z(z)} - \frac{\bar{P}_Z(z)}{P_Z(z)} \right)^2 \right] \quad (214)$$

$$\begin{aligned} &= \mathbb{E}_{P_Z} \mathbb{E}_{\mathcal{C}} \left[\frac{1}{|\mathcal{I}|^2} \sum_{i \in \mathcal{I}} \left(\frac{\bar{P}_{Z|U}(z|u_i)}{P_Z(z)} \right)^2 + \sum_{i \in \mathcal{I}} \sum_{j \neq i} \frac{1}{|\mathcal{I}|^2} \frac{\bar{P}_{Z|U}(z|u_i)}{P_Z(z)} \frac{\bar{P}_{Z|U}(z|u_j)}{P_Z(z)} + \dots \right. \\ & \quad \left. - \sum_{i \in \mathcal{I}} \frac{2}{|\mathcal{I}|} \frac{\bar{P}_{Z|U}(z|u_i)}{P_Z(z)} \frac{\bar{P}_Z(z)}{P_Z(z)} + \left(\frac{\bar{P}_Z(z)}{P_Z(z)} \right)^2 \right] \end{aligned} \quad (215)$$

$$= \mathbb{E}_{P_Z} \frac{1}{|\mathcal{I}|} \mathbb{E}_{P_U} \left[\left(\frac{\bar{P}_{Z|U}(z|u)}{P_Z(z)} \right)^2 - \left(\frac{\bar{P}_Z(z)}{P_Z(z)} \right)^2 \right] \quad (216)$$

$$\leq \frac{1}{|\mathcal{I}|} \mathbb{E}_{P_Z} \mathbb{E}_{P_U} \left[\left(\frac{\bar{P}_{Z|U}(z|u)}{P_Z(z)} \right)^2 \right] \quad (217)$$

$$= \frac{1}{|\mathcal{I}|} \sum_{u, z} \frac{P_U(u) \bar{P}_{Z|U}(z|u)^2}{P_Z(z)} \mathbf{1}[(u, z) \in \mathcal{T}_c(\gamma_c)] \quad (218)$$

$$= \frac{1}{|\mathcal{I}|} \sum_{(u, z) \in \mathcal{T}_c(\gamma_c)} \frac{P_U(u) P_{Z|U}(z|u)^2}{P_Z(z)} \quad (219)$$

$$= \frac{1}{|\mathcal{I}|} \Delta(\gamma_c, P_{UZ}) \quad (220)$$

where (216) follows from the fact that for $i \neq j$,

$$\mathbb{E}_{\mathcal{C}} \left[\frac{\bar{P}_{Z|U}(z|u_i)}{P_Z(z)} \right] = \frac{\bar{P}_Z(z)}{P_Z(z)}, \quad (221)$$

$$\mathbb{E}_{\mathcal{C}} \left[\frac{\bar{P}_{Z|U}(z|u_i)}{P_Z(z)} \frac{\bar{P}_{Z|U}(z|u_j)}{P_Z(z)} \right] = \mathbb{E}_{\mathcal{C}} \left[\frac{\bar{P}_{Z|U}(z|u_i)}{P_Z(z)} \right] \mathbb{E}_{\mathcal{C}} \left[\frac{\bar{P}_{Z|U}(z|u_j)}{P_Z(z)} \right] = \left(\frac{\bar{P}_Z(z)}{P_Z(z)} \right)^2, \quad (222)$$

and (219) follows from the definition of $\bar{P}_{Z|U}(z|u)$ in (204). Uniting (208) and (220) completes the proof of Lemma 34. ■

We can relax (207) as

$$\mathbb{E}_{\mathcal{C}} [d(P_{\tilde{Z}}, P_Z)] \leq P_{UZ} (\mathcal{T}_c(\gamma_c)^c) + \frac{1}{2} \sqrt{\frac{2^{\gamma_c}}{|\mathcal{I}|}} \quad (223)$$

by upper bounding $\Delta(\gamma_c, P_{UZ})$; cf. (75).

In some cases, we need to consider the noiseless channel, i.e., $\mathcal{Z} = \mathcal{U}$ and $W(z|u) = \mathbf{1}[u = z]$, and want to approximate P_U by

$$P_{\tilde{U}}(u) = \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \mathbf{1}[u_i = u]. \quad (224)$$

In this case, the bound (223) reduces to

$$\mathbb{E}_{\mathcal{C}} [d(\bar{P}_{\tilde{U}}, \bar{P}_U)] \leq P_U [-\log P_U(u) > \gamma_c] + \frac{1}{2} \sqrt{\frac{2^{\gamma_c}}{|\mathcal{I}|}}. \quad (225)$$

APPENDIX C SIMULATION OF TEST CHANNEL

In this appendix, we develop two lemmas which form crucial components of the proof of all CS-type bounds. To do this, we consider the problem of channel simulation [15]–[19]. Roughly speaking, the problem is described as follows. Let a joint distribution P_{UZ} on $\mathcal{U} \times \mathcal{Z}$ given. An observer (encoder) of Z describes to a distant random number generator (decoder) that produces \tilde{U} so that simulated channel $P_{\tilde{U}|Z}$ is statistically indistinguishable from $P_{U|Z}$. We assume that the encoder and decoder have common randomness.

In the following, we construct a pair of encoder and decoder for this problem. More precisely, we construct a stochastic map from $\mathcal{K} \times \mathcal{Z}$ to \mathcal{L} and a map from $\mathcal{K} \times \mathcal{L}$ to \mathcal{U} , where \mathcal{K} is the alphabet of the common randomness and \mathcal{L} is a message index set. We will use notations introduced in Appendix B.

To construct a stochastic map from $\mathcal{K} \times \mathcal{Z}$ to \mathcal{L} , we first consider the channel resolvability code as follows. Let us generate a codebook $\mathcal{C} = \{u_{11}, \dots, u_{|\mathcal{K}||\mathcal{L}|}\}$, where each codeword u_{kl} is randomly and independently generated from P_U , which is the marginal of P_{UZ} . Let K and L be the uniform random numbers on \mathcal{K} and \mathcal{L} respectively. Moreover, let $\bar{P}_{Z|U}$ be a smoothed version of $P_{Z|U}$ defined in (204). Then, \mathcal{C} , K , L , and $\bar{P}_{Z|U}$ induce the sub-normalized measure

$$\bar{P}_{KL\tilde{U}\tilde{Z}}(k, l, u, z) := \frac{1}{|\mathcal{K}||\mathcal{L}|} \bar{P}_{Z|U}(z|u) \mathbf{1}[u_{kl} = u]. \quad (226)$$

Marginals $\bar{P}_{\tilde{U}\tilde{Z}|K}$, $\bar{P}_{L\tilde{Z}|K}$, and $\bar{P}_{\tilde{Z}|K}$ of $\bar{P}_{KL\tilde{U}\tilde{Z}}$ are also induced as

$$\bar{P}_{\tilde{U}\tilde{Z}|K}(u, z|k) = \sum_{l \in \mathcal{L}} \frac{1}{|\mathcal{L}|} \bar{P}_{Z|U}(z|u) \mathbf{1}[u_{kl} = u] \quad (227)$$

$$\bar{P}_{L\tilde{Z}|K}(l, z|k) = \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{L}|} \bar{P}_{Z|U}(z|u) \mathbf{1}[u_{kl} = u] \quad (228)$$

and

$$\bar{P}_{\tilde{Z}|K}(z|k) = \sum_{u, l} \frac{1}{|\mathcal{L}|} \bar{P}_{Z|U}(z|u) \mathbf{1}[u_{kl} = u]. \quad (229)$$

Now, we define a stochastic map $\varphi_{\mathcal{C}}: \mathcal{K} \times \mathcal{Z} \rightarrow \mathcal{L}$ as⁶

$$\varphi_{\mathcal{C}}(l|k, z) = \frac{\bar{P}_{L\tilde{Z}|K}(l, z|k)}{\bar{P}_{\tilde{Z}|K}(z|k)}. \quad (230)$$

⁶When $\bar{P}_{\tilde{Z}|K}(z|k) = 0$, we define $\varphi_{\mathcal{C}}(l|k, z)$ arbitrarily.

By using \mathcal{C} and $\varphi_{\mathcal{C}}$, we can simulate the channel $P_{U|Z}$ as follows. The observer of Z (encoder) sends the realization $l \in \mathcal{L}$ of the random experiment $\varphi_{\mathcal{C}}(\cdot|k, z)$ if the common randomness is k and $Z = z$. Receiving the index $l \in \mathcal{L}$ from the observer and the common randomness $k \in \mathcal{K}$, the random number generator (decoder) outputs $u_{kl} \in \mathcal{C}$.

Let \hat{L} be the output of the stochastic map $\varphi_{\mathcal{C}}$ for the inputs K and Z . Then, the output \hat{U} of the decoder is $\hat{U} = u_{K\hat{L}}$ and the joint distribution of K, \hat{L}, \hat{U}, Z is given by

$$P_{K\hat{L}\hat{U}Z}(k, l, u, z) = \frac{1}{|\mathcal{K}|} P_Z(z) \varphi_{\mathcal{C}}(l|k, z) \mathbf{1}[u_{kl} = u]. \quad (231)$$

It should be noted here that the conditional distribution $P_{\hat{U}|KZ}$ induced by $P_{K\hat{L}\hat{U}Z}$ satisfies

$$P_{\hat{U}|KZ}(u|k, z) = \frac{\bar{P}_{\hat{U}\hat{Z}|K}(u, z|k)}{\bar{P}_{\hat{Z}|K}(z|k)} = \sum_{l \in \mathcal{L}} \varphi_{\mathcal{C}}(l|k, z) \mathbf{1}[u_{kl} = u]. \quad (232)$$

We also introduce a smoothed version of $P_{K\hat{L}\hat{U}Z}$ as follows:

$$\bar{P}_{K\hat{L}\hat{U}Z}(k, l, u, z) = \frac{1}{|\mathcal{K}|} \bar{P}_Z(z) \varphi_{\mathcal{C}}(l|k, z) \mathbf{1}[u_{kl} = u] \quad (233)$$

where \bar{P}_Z is the marginal of $\bar{P}_{UZ} := P_U \bar{P}_{Z|U}$; i.e. $\bar{P}_Z(z) := \sum_u P_U(u) \bar{P}_{Z|U}(z|u)$.

Now, we prove two lemmas which can be used to evaluate the performance of the channel simulation model described above.

Lemma 35. *We have*

$$d(P_{\hat{U}Z}, \bar{P}_{UZ}) \leq \frac{P_{UZ}((u, z) \notin \mathcal{T}_c(\gamma_c))}{2} + d(\bar{P}_{\hat{U}Z}, \bar{P}_{UZ}) \quad (234)$$

where $P_{\hat{U}Z}$ (resp. $\bar{P}_{\hat{U}Z}$) is the marginal of $P_{K\hat{L}\hat{U}Z}$ (resp. $\bar{P}_{K\hat{L}\hat{U}Z}$).

Proof: By the triangular inequality, we have

$$d(P_{\hat{U}Z}, \bar{P}_{UZ}) \leq d(P_{\hat{U}Z}, \bar{P}_{\hat{U}Z}) + d(\bar{P}_{\hat{U}Z}, \bar{P}_{UZ}). \quad (235)$$

Further, we can bound the first term of the right hand side of the above inequality as

$$d(P_{\hat{U}Z}, \bar{P}_{\hat{U}Z}) = d(P_Z P_{\hat{U}|Z}, \bar{P}_Z \bar{P}_{\hat{U}|Z}) \quad (236)$$

$$\leq d(P_Z, \bar{P}_Z) \quad (237)$$

$$\leq d(P_{UZ}, \bar{P}_{UZ}) \quad (238)$$

$$= \frac{P_{UZ}((u, z) \in \mathcal{T}_c(\gamma_c)^c)}{2} \quad (239)$$

where (237) follows from the fact that

$$P_{\hat{U}|Z}(u|z) = \bar{P}_{\hat{U}|Z}(u|z) = \sum_{k, l} \frac{1}{|\mathcal{K}|} \varphi_{\mathcal{C}}(l|k, z) \mathbf{1}[u_{kl} = u] \quad (240)$$

and the data-processing inequality (182), (238) follows from the monotonicity property in (181), and (239) follows from (200). \blacksquare

Lemma 36. *For every $\gamma > 0$, we have*

$$\mathbb{E}_{\mathcal{C}}[d(\bar{P}_{\hat{U}Z}, \bar{P}_{UZ})] \leq \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UZ})}{|\mathcal{L}|}} + \frac{1}{2} \sqrt{\frac{2\gamma}{|\mathcal{K}||\mathcal{L}|}} + P_U[-\log P_U(u) > \gamma]. \quad (241)$$

Proof: We have

$$d(\bar{P}_{\hat{U}Z}, \bar{P}_{UZ}) \leq d(\bar{P}_{\hat{U}Z}, \bar{P}_{\hat{U}\hat{Z}}) + d(\bar{P}_{\hat{U}\hat{Z}}, \bar{P}_{UZ}) \quad (242)$$

$$\leq d(\bar{P}_{K\hat{U}Z}, \bar{P}_{K\hat{U}\hat{Z}}) + d(\bar{P}_{\hat{U}\hat{Z}}, \bar{P}_{UZ}) \quad (243)$$

$$\leq d(\bar{P}_{K\hat{U}Z}, \bar{P}_{K\hat{U}\hat{Z}}) + d(P_{\hat{U}}, P_U) \quad (244)$$

where (242) follows from the triangle inequality, (243) follows from the monotonicity (181), and (244) follows from the data-processing inequality (182) and the fact that, by the definition of $\bar{P}_{KL\tilde{U}\tilde{Z}}$,

$$\bar{P}_{\tilde{Z}|\tilde{U}}(z|u) = \bar{P}_{Z|U}(z|u). \quad (245)$$

We bound the first term in (244) as follows:

$$d(\bar{P}_{K\hat{U}Z}, \bar{P}_{K\tilde{U}\tilde{Z}}) = d(P_K \bar{P}_Z \bar{P}_{\hat{U}|KZ}, \bar{P}_{K\tilde{U}\tilde{Z}}) \quad (246)$$

$$= d(P_K \bar{P}_Z \bar{P}_{\hat{U}|KZ}, P_K \bar{P}_{\tilde{U}|\tilde{Z}|K}) \quad (247)$$

$$= d(P_K \bar{P}_Z \bar{P}_{\hat{U}|KZ}, P_K \bar{P}_{\tilde{Z}|K} P_{\hat{U}|KZ}) \quad (248)$$

$$\leq d(P_K \bar{P}_Z, P_K \bar{P}_{\tilde{Z}|K}) \quad (249)$$

$$\leq \sum_{k \in \mathcal{K}} \frac{1}{|\mathcal{K}|} d(\bar{P}_Z(\cdot), \bar{P}_{\tilde{Z}|K}(\cdot|k)) \quad (250)$$

where (248) follows from the first equality in (232), and (249) follows from the data-processing inequality (182) and the fact that

$$P_{\hat{U}|KZ}(u|k, z) = \bar{P}_{\hat{U}|KZ}(u|k, z) = \sum_l \varphi_{\mathcal{C}}(l|k, z) \mathbf{1}[u_{kl} = u]. \quad (251)$$

Taking the expectation with respect to the codebook \mathcal{C} yields

$$\mathbb{E}_{\mathcal{C}} [d(\bar{P}_{K\hat{U}Z}, \bar{P}_{K\tilde{U}\tilde{Z}})] \leq \sum_k \frac{1}{|\mathcal{K}|} \mathbb{E}_{\mathcal{C}} [d(\bar{P}_Z(\cdot), \bar{P}_{\tilde{Z}|K}(\cdot|k))] \quad (252)$$

$$\leq \frac{1}{2} \sqrt{\frac{\Delta(\gamma_{\mathcal{C}}, P_{UZ})}{|\mathcal{L}|}}. \quad (253)$$

where the second inequality is obtained by using (206) for each $k \in \mathcal{K}$.

Further, we have

$$\mathbb{E}_{\mathcal{C}} [d(P_{\tilde{U}}, P_U)] \leq P_U [-\log P_U(u) > \gamma] + \frac{1}{2} \sqrt{\frac{2^\gamma}{|\mathcal{K}||\mathcal{L}|}}. \quad (254)$$

by (225).

Combining (244), (253), and (254) completes the proof of the lemma. \blacksquare

APPENDIX D

PROOF OF THE FIRST NON-ASYMPTOTIC BOUND FOR WAK IN THEOREM 14

A. Code Construction

We construct a WAK code by using the common randomness for the helper and the decoder. To do this, we use the stochastic map introduced in Appendix C. Let \mathcal{K} be the alphabet of the common randomness for the helper and the decoder. Further, let $\mathcal{Z} = \mathcal{Y}$ and $Z = Y$, that is, let $P_{UZ} = P_{UY}$, where P_{UY} is the marginal of the given distribution $P_{UXY} \in \mathcal{P}(P_{XY})$. It should be noted here that, in this case, $\mathcal{T}_{\mathcal{C}}(\gamma_{\mathcal{C}})$ defined in (203) is equivalent to $\mathcal{T}_{\mathcal{C}}^{\text{WAK}}(\gamma_{\mathcal{C}})$ defined in (40). Now, let us consider the stochastic map $\varphi_{\mathcal{C}}$ defined in (230).

By using $\varphi_{\mathcal{C}}$, we construct a WAK code Φ as follows. The main encoder uses a random bin coding $f: \mathcal{X} \rightarrow \mathcal{M}$. The helper uses the stochastic map $\varphi_{\mathcal{C}}: \mathcal{K} \times \mathcal{Y} \rightarrow \mathcal{L}$. That is, when the side information is $y \in \mathcal{Y}$ and the common randomness is $k \in \mathcal{K}$, the helper generates $l \in \mathcal{L}$ according to $\varphi_{\mathcal{C}}(\cdot|k, y)$ and sends l to the decoder. For given $m \in \mathcal{M}$, $l \in \mathcal{L}$, and common randomness $k \in \mathcal{K}$, the decoder outputs the unique $\hat{x} \in \mathcal{X}$ such that $f(\hat{x}) = m$ and

$$(u_{kl}, \hat{x}) \in \mathcal{T}_{\mathbf{b}}^{\text{WAK}}(\gamma_{\mathbf{b}}). \quad (255)$$

If no such unique \hat{x} exists, or if there is more than one such \hat{x} , then a decoding error is declared.

B. Analysis of Error Probability

Let \hat{L} be the random index chosen by the helper via the stochastic map $\varphi_{\mathcal{C}}(\cdot|K, Y)$, and let $\hat{U} = u_{K\hat{L}}$. Note that the joint distribution of K, \hat{L}, \hat{U}, Y is given as follows; cf. (231)

$$P_{K\hat{L}\hat{U}Y}(k, l, u, y) = \frac{1}{|\mathcal{K}|} P_Y(y) \varphi_{\mathcal{C}}(l|k, y) \mathbf{1}[u_{kl} = u] \quad (256)$$

and then, the joint distribution of K, \hat{L}, \hat{U}, Y and X is given as

$$P_{K\hat{L}\hat{U}XY}(k, l, u, x, y) = P_{K\hat{L}\hat{U}Y}(k, l, u, y) P_{X|Y}(x|y). \quad (257)$$

The smoothed versions $\bar{P}_{K\hat{L}\hat{U}Y}$ and $\bar{P}_{K\hat{L}\hat{U}XY}$ are given by substituting P_Y in (256) with \bar{P}_Y ; cf. (233).

If the decoding error occurs, at least one of the following events occurs:

$$\begin{aligned} \mathcal{E}_1 &:= \{(u_{kl}, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b)\} \\ \mathcal{E}_2 &:= \{\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u_{kl}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)\} \end{aligned}$$

Hence, the error probability averaged over random coding f , the random codebook \mathcal{C} , and the common randomness K can be bounded as

$$\mathbb{E}_f \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbb{P}_e(\Phi)] = \mathbb{E}_f \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}XY}(\mathcal{E}_1 \cup \mathcal{E}_2)]. \quad (258)$$

Let

$$\mathcal{E}_{12} := \{(u, x) : (u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \text{ or } \exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b(\gamma_b)\}. \quad (259)$$

Then, for fixed f and \mathcal{C} , we have

$$\begin{aligned} &P_{K\hat{L}\hat{U}XY}(\mathcal{E}_1 \cup \mathcal{E}_2) \\ &= \sum_{k, l, u, x, y} P_{K\hat{L}\hat{U}XY}(k, l, u, x, y) \mathbf{1}[(u, x) \in \mathcal{E}_{12}] \end{aligned} \quad (260)$$

$$= \sum_{k, l, u, x, y} P_{K\hat{L}\hat{U}Y}(k, l, u, y) P_{X|Y}(x|y) \mathbf{1}[(u, x) \in \mathcal{E}_{12}] \quad (261)$$

$$= \sum_{u, x, y} P_{\hat{U}Y}(u, y) P_{X|Y}(x|y) \mathbf{1}[(u, x) \in \mathcal{E}_{12}] \quad (262)$$

$$= P_{\hat{U}XY}((u, x) \in \mathcal{E}_{12}) \quad (263)$$

$$\leq \bar{P}_{UXY}((u, x) \in \mathcal{E}_{12}) + \frac{1 - \bar{P}_{UXY}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})}{2} + d(P_{\hat{U}XY}, \bar{P}_{UXY}) \quad (264)$$

$$= \bar{P}_{UXY}((u, x) \in \mathcal{E}_{12}) + \frac{P_{UXY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c))}{2} + d(P_{\hat{U}XY}, \bar{P}_{UXY}) \quad (265)$$

$$= \bar{P}_{UXY}((u, x) \in \mathcal{E}_{12}) + \frac{P_{UY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c))}{2} + d(P_{\hat{U}Y} P_{X|Y}, \bar{P}_{UY} P_{X|Y}) \quad (266)$$

$$\leq \bar{P}_{UXY}((u, x) \in \mathcal{E}_{12}) + \frac{P_{UY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c))}{2} + d(P_{\hat{U}Y}, \bar{P}_{UY}) \quad (267)$$

$$\leq \bar{P}_{UXY}((u, x) \in \mathcal{E}_{12}) + P_{UY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) + d(\bar{P}_{\hat{U}Y}, \bar{P}_{UY}) \quad (268)$$

$$\begin{aligned} &\leq \bar{P}_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b)) + \bar{P}_{UXY}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b(\gamma_b)] \\ &\quad + P_{UY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) + d(\bar{P}_{\hat{U}Y}, \bar{P}_{UY}) \end{aligned} \quad (269)$$

$$\begin{aligned} &= P_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cap (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)) + P_{UY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) + d(\bar{P}_{\hat{U}Y}, \bar{P}_{UY}) \\ &\quad + \bar{P}_{UXY}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b(\gamma_b)] \end{aligned} \quad (270)$$

$$\begin{aligned} &= P_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cup (u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) + d(\bar{P}_{\hat{U}Y}, \bar{P}_{UY}) \\ &\quad + \bar{P}_{UXY}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b(\gamma_b)] \end{aligned} \quad (271)$$

where (264) follows from (183), (267) follows from the data-processing inequality (182), and (268) follows from Lemma 35. By taking average over \mathcal{C} , the second term in (271) is upper bounded

$$\mathbb{E}_{\mathcal{C}} [d(\bar{P}_{\hat{U}Y}, \bar{P}_{UY})] \leq \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UY})}{|\mathcal{L}|}} + \frac{1}{2} \sqrt{\frac{2^\gamma}{|\mathcal{K}||\mathcal{L}|}} + P_U [-\log P_U(u) > \gamma] \quad (272)$$

by using Lemma 36. On the other hand, by taking average over f , the last term in (271) is upper bounded as

$$\begin{aligned} & \mathbb{E}_f [\bar{P}_{UXY}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b(\gamma_b)]] \\ & \leq \sum_{u,x,y} \bar{P}_{UXY}(u, x, y) \sum_{\tilde{x} \neq x} \mathbb{E}_f [\mathbf{1}[f(\tilde{x}) = f(x)] \mathbf{1}[(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)]] \end{aligned} \quad (273)$$

$$\leq \frac{1}{|\mathcal{M}|} \sum_u P_U(u) \sum_{\tilde{x}} \mathbf{1}[(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \quad (274)$$

$$= \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) \quad (275)$$

where we used the fact $\sum_{x,y} \bar{P}_{UXY}(u, x, y) \leq P_U(u)$ in (274). Hence, by (271), (272), and (275), we have

$$\begin{aligned} \mathbb{E}_f \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbf{P}_e(\Phi)] &= \mathbb{E}_f \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}XY}(\mathcal{E}_1 \cup \mathcal{E}_2)] \\ &\leq P_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cup (u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) \end{aligned} \quad (276)$$

$$\begin{aligned} &+ \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UY})}{|\mathcal{L}|}} + \frac{1}{2} \sqrt{\frac{2^\gamma}{|\mathcal{K}||\mathcal{L}|}} + P_U [-\log P_U(u) > \gamma] \\ &+ \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) \end{aligned} \quad (277)$$

Since we can choose $\gamma > 0$ and $|\mathcal{K}|$ arbitrarily large, we have

$$\begin{aligned} & \mathbb{E}_f \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbf{P}_e(\Phi)] \\ & \leq P_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cup (u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) \\ & \quad + \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UY})}{|\mathcal{L}|}} + \delta. \end{aligned} \quad (278)$$

Consequently, there exists at least one code (k, f, \mathcal{C}) such that $\mathbf{P}_e(\Phi)$ is smaller than the right-hand-side of the inequality above. This completes the proof of Theorem 14.

APPENDIX E

PROOF OF THE SECOND NON-ASYMPTOTIC BOUND FOR WAK IN THEOREM 16

To prove Theorem 16, we modify the proof of Theorem 14 as follows.

First, we use $\mathcal{J} = \{1, \dots, J\}$ instead of \mathcal{L} in the construction of $\varphi_{\mathcal{C}}$, where J is the given integer.

Then, the helper and the decoder are modified as follows. The helper first uses the stochastic map $\varphi_{\mathcal{C}}: \mathcal{K} \times \mathcal{Y} \rightarrow \mathcal{J}$. That is, it generates $j \in \mathcal{J}$ according to $\varphi_{\mathcal{C}}(\cdot | k, y)$ when the side information is $y \in \mathcal{Y}$ and the common randomness is $k \in \mathcal{K}$. Then, the helper sends j by using random bin coding $\kappa: \mathcal{J} \rightarrow \mathcal{L}$. This means that to every $j \in \mathcal{J}$, it independently and uniformly assigns a random index $l \in \mathcal{L}$. For given $m \in \mathcal{M}$, $l \in \mathcal{L}$, and the common randomness $k \in \mathcal{K}$, the decoder outputs the unique $\hat{x} \in \mathcal{X}$ such that $f(\hat{x}) = m$ and

$$(u_{kj}, \hat{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b) \quad (279)$$

for some $j \in \mathcal{J}$ satisfying $\kappa(j) = l$. If no such unique \hat{x} exists, or if there is more than one such \hat{x} , then a decoding error is declared.

The analysis of the probability of the decoding error is modified as follows: Let \hat{J} be the random index chosen by the helper via the stochastic map $\varphi_{\mathcal{C}}(\cdot|K, Y)$, and let $\hat{U} = u_{K\hat{J}}$. The joint distribution of K, \hat{J}, \hat{U}, Y is

$$P_{K\hat{J}\hat{U}Y}(k, j, u, y) = \frac{1}{|\mathcal{K}|} P_Y(y) \varphi_{\mathcal{C}}(j|k, y) \mathbf{1}[u_{kj} = u]. \quad (280)$$

The other measures $P_{K\hat{J}\hat{U}XY}$, $\bar{P}_{K\hat{J}\hat{U}Y}$, and $\bar{P}_{K\hat{J}\hat{U}XY}$ are given similarly. On the other hand, if the decoding error occurs, at least one of the following occurs:

$$\mathcal{E}_1 := \{(u_{kj}, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b)\} \quad (281)$$

$$\mathcal{E}_2 := \{\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u_{kj}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)\} \quad (282)$$

$$\mathcal{E}_3 := \left\{ \exists \tilde{x} \neq x, \tilde{j} \neq j \text{ s.t. } f(\tilde{x}) = f(x), \kappa(\tilde{j}) = \kappa(j), (u_{k\tilde{j}}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b) \right\}. \quad (283)$$

Hence, the error probability averaged over random coding f, κ , the random codebook \mathcal{C} , and the common randomness K can be bounded as

$$\mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbb{P}_e(\Phi)] \leq \mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} [P_{K\hat{J}\hat{U}XY}(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3)] \quad (284)$$

$$\leq \mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} [P_{K\hat{J}\hat{U}XY}(\mathcal{E}_1 \cup \mathcal{E}_2)] + \mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} [P_{K\hat{J}\hat{U}XY}(\mathcal{E}_3)]. \quad (285)$$

The first term in (285) is upper bounded in the same way as bounding (258), and we have

$$\begin{aligned} \mathbb{E}_f \mathbb{E}_{\mathcal{C}} [P_{K\hat{J}\hat{U}XY}(\mathcal{E}_1 \cup \mathcal{E}_2)] &\leq P_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cup (u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) \\ &\quad + \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UY})}{|\mathcal{J}|}} + \delta. \end{aligned} \quad (286)$$

On the other hand, the second term in (285) is upper bounded as follows.

$$\begin{aligned} &\mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} [P_{K\hat{J}\hat{U}XY}(\mathcal{E}_3)] \\ &= \mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} \left[\sum_{k, l, u, x, y} P_{K\hat{J}\hat{U}XY}(k, j, u, x, y) \mathbf{1}[\exists \tilde{x} \neq x, \tilde{j} \neq j \text{ s.t. } f(\tilde{x}) = f(x), \kappa(\tilde{j}) = \kappa(j), (u_{k\tilde{j}}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \end{aligned} \quad (287)$$

$$\leq \mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} \left[\sum_{k, j, u, x, y} P_{K\hat{J}\hat{U}XY}(k, j, u, x, y) \sum_{\substack{\tilde{x} \neq x \\ \tilde{j} \neq j}} \mathbf{1}[f(\tilde{x}) = f(x), \kappa(\tilde{j}) = \kappa(j)] \cdot \mathbf{1}[(u_{k\tilde{j}}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \quad (288)$$

$$= \mathbb{E}_{f, \kappa} \mathbb{E}_{\mathcal{C}} \left[\sum_{k, j, x, y} P_{K\hat{J}XY}(k, j, x, y) \sum_{\substack{\tilde{x} \neq x \\ \tilde{j} \neq j}} \mathbf{1}[f(\tilde{x}) = f(x), \kappa(\tilde{j}) = \kappa(j)] \cdot \mathbf{1}[(u_{k\tilde{j}}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \quad (289)$$

$$\leq \frac{1}{|\mathcal{M}||\mathcal{L}|} \mathbb{E}_{\mathcal{C}} \left[\sum_{k, j, x, y} P_{K\hat{J}XY}(k, j, x, y) \sum_{\tilde{x}, \tilde{j}} \mathbf{1}[(u_{k\tilde{j}}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \quad (290)$$

$$= \frac{1}{|\mathcal{M}||\mathcal{L}|} \mathbb{E}_{\mathcal{C}} \left[\sum_k P_K(k) \sum_{\tilde{x}, \tilde{j}} \mathbf{1}[(u_{k\tilde{j}}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \quad (291)$$

$$= \frac{1}{|\mathcal{M}||\mathcal{L}|} \mathbb{E}_{\mathcal{C}} \left[\sum_k \frac{1}{|\mathcal{K}|} \sum_{\tilde{x}, \tilde{j}} \mathbf{1}[(u_{k\tilde{j}}, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \quad (292)$$

$$= \frac{|\mathcal{J}|}{|\mathcal{M}||\mathcal{L}|} \sum_{u, \tilde{x}} P_U(u) \mathbf{1}[(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \quad (293)$$

$$= \frac{|\mathcal{J}|}{|\mathcal{M}||\mathcal{L}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) \quad (294)$$

where (293) follows from the fact that \mathcal{C} is generated according to P_U . Substituting (286) and (294) into (285), we have Theorem 16.

APPENDIX F PROOF OF THE NON-ASYMPTOTIC BOUND FOR WZ IN THEOREM 17

A. Code Construction

Similar to WAK coding in the previous two sections, we use the stochastic map introduced in Appendix C. In WZ coding, let \mathcal{K} be the alphabet of the common randomness for the encoder and the decoder, and let $\mathcal{Z} = \mathcal{X}$ and $P_{UZ} = P_{UX}$. Note that $\mathcal{T}_c(\gamma_c)$ defined in (203) is equivalent to $\mathcal{T}_c^{\text{WZ}}(\gamma_c)$ defined in (54). Now, let us consider the stochastic map $\varphi_{\mathcal{C}}$ defined in (230).

By using $\varphi_{\mathcal{C}}$, we construct a WZ code Φ as follows. The encoder first uses the stochastic map $\varphi_{\mathcal{C}}: \mathcal{K} \times \mathcal{X} \rightarrow \mathcal{L}$. That is, it generates $l \in \mathcal{L}$ according to $\varphi_{\mathcal{C}}(\cdot | k, y)$ when the source output is $x \in \mathcal{X}$ and the common randomness is $k \in \mathcal{K}$. Then, the encoder sends l by using random bin coding $\kappa: \mathcal{L} \rightarrow \mathcal{M}$. This means that to every $l \in \mathcal{L}$, it independently and uniformly assigns a random index $m \in \mathcal{M}$. For given $m \in \mathcal{M}$, $y \in \mathcal{Y}$, and the common randomness $k \in \mathcal{K}$, the decoder finds the unique index $l \in \mathcal{L}$ such that $\kappa(l) = m$ and

$$(u_{kl}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p). \quad (295)$$

Then, decoder outputs $g(u_{kl}, y) \in \hat{\mathcal{X}}$. If no unique l satisfying (295) exists, or if there is more than one such l satisfying (295), then a decoding error is declared.

B. Analysis of Probability of Excess Distortion

Let \hat{L} be the random index chosen by the encoder via the stochastic map $\varphi_{\mathcal{C}}(\cdot | K, X)$, and let $\hat{U} = u_{K\hat{L}}$. Note that the joint distribution of K, \hat{L}, \hat{U}, X is given as follows; cf. (231)

$$P_{K\hat{L}\hat{U}X}(k, l, u, x) = \frac{1}{|\mathcal{K}|} P_X(x) \varphi_{\mathcal{C}}(l | k, x) \mathbf{1}[u_{kl} = u] \quad (296)$$

and then, the joint distribution of K, \hat{L}, \hat{U}, X and Y is given as

$$P_{K\hat{L}\hat{U}XY}(k, l, u, x, y) = P_{K\hat{L}\hat{U}X}(k, l, u, x) P_{Y|X}(y | x). \quad (297)$$

The smoothed versions $\bar{P}_{K\hat{L}\hat{U}X}$ and $\bar{P}_{K\hat{L}\hat{U}XY}$ are given by substituting P_X in (296) with \bar{P}_X ; cf. (233).

If the distortion exceeds D , at least one of the following events occurs:

$$\mathcal{E}_0 := \{(u_{kl}, x, y) \notin \mathcal{T}_d^{\text{WZ}}(D)\} \quad (298)$$

$$\mathcal{E}_1 := \{(u_{kl}, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)\} \quad (299)$$

$$\mathcal{E}_2 := \left\{ \exists \tilde{l} \neq l \text{ s.t. } \kappa(\tilde{l}) = \kappa(l), (u_{k\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p) \right\}. \quad (300)$$

Hence, the probability of excess distortion averaged over the random coding κ , the random codebook \mathcal{C} , and the common randomness K can be bounded as

$$\mathbb{E}_{\kappa} \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbb{P}_e(\Phi; D)] \leq \mathbb{E}_{\kappa} \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}XY}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2)] \quad (301)$$

$$\leq \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}XY}(\mathcal{E}_0 \cup \mathcal{E}_1)] + \mathbb{E}_{\kappa} \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}XY}(\mathcal{E}_2)]. \quad (302)$$

At first, we evaluate the first term in (302). For fixed \mathcal{C} ,

$$P_{K\hat{L}\hat{U}XY}(\mathcal{E}_0 \cup \mathcal{E}_1) = \sum_{k,l,u,x,y} P_{K\hat{L}\hat{U}X}(k,l,u,x) P_{Y|X}(y|x) \mathbf{1}[(u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \quad (303)$$

$$= \sum_{u,x,y} P_{\hat{U}X}(u,x) P_{Y|X}(y|x) \mathbf{1}[(u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \quad (304)$$

$$= P_{\hat{U}XY}((u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) \quad (305)$$

$$\leq \bar{P}_{UXY}((u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) + \frac{1 - \bar{P}_{UXY}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})}{2} + d(P_{\hat{U}XY}, \bar{P}_{UXY}) \quad (306)$$

$$= \bar{P}_{UXY}((u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) + \frac{P_{UXY}((u,x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c))}{2} + d(P_{\hat{U}XY}, \bar{P}_{UXY}) \quad (307)$$

$$= \bar{P}_{UXY}((u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) + \frac{P_{UX}((u,x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c))}{2} + d(P_{\hat{U}X}, \bar{P}_{UX}) \quad (308)$$

$$\leq \bar{P}_{UXY}((u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) + \frac{P_{UX}((u,x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c))}{2} + d(P_{\hat{U}X}, \bar{P}_{UX}) \quad (309)$$

$$\leq \bar{P}_{UXY}((u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) + P_{UX}((u,x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c)) + d(\bar{P}_{\hat{U}X}, \bar{P}_{UX}) \quad (310)$$

$$= P_{UXY}((u,x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c) \cup (u,x,y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u,y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) + d(\bar{P}_{\hat{U}X}, \bar{P}_{UX}) \quad (311)$$

where (306) follows from (183), (309) follows from the data-processing inequality (182), (310) follows from Lemma 35, and (311) follows from the fact that \bar{P}_{UXY} is the smoothed version of P_{UXY} with respect to $\mathcal{T}_c^{\text{WZ}}(\gamma_c)$. By taking the average over \mathcal{C} , we see that the second term in (311) is upper bounded as

$$\mathbb{E}_{\mathcal{C}} [d(\bar{P}_{\hat{U}X}, \bar{P}_{UX})] \leq \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UX})}{|\mathcal{L}|}} + \frac{1}{2} \sqrt{\frac{2\gamma}{|\mathcal{K}||\mathcal{L}|}} + P_U [-\log P_U(u) > \gamma] \quad (312)$$

where we used Lemma 36.

Next, we upper bound the second term in (302):

$$\mathbb{E}_{\kappa} \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}XY}(\mathcal{E}_2)] = \mathbb{E}_{\kappa} \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,u,x,y} P_{K\hat{L}\hat{U}XY}(k,l,u,x,y) \mathbf{1}[\exists \tilde{l} \neq l \text{ s.t. } \kappa(\tilde{l}) = \kappa(l), (u_{k\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \quad (313)$$

$$\leq \mathbb{E}_{\kappa} \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,u,x,y} P_{K\hat{L}\hat{U}XY}(k,l,u,x,y) \sum_{\tilde{l} \neq l} \mathbf{1}[\kappa(\tilde{l}) = \kappa(l)] \cdot \mathbf{1}[(u_{k\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \quad (314)$$

$$= \mathbb{E}_{\kappa} \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,x,y} P_{K\hat{L}XY}(k,l,x,y) \sum_{\tilde{l} \neq l} \mathbf{1}[\kappa(\tilde{l}) = \kappa(l)] \cdot \mathbf{1}[(u_{k\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \quad (315)$$

$$\leq \frac{1}{|\mathcal{M}|} \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,x,y} P_{K\hat{L}XY}(k,l,x,y) \sum_{\tilde{l}} \mathbf{1}[(u_{k\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \quad (316)$$

$$= \frac{1}{|\mathcal{M}|} \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,x,y} P_{K\hat{L}X}(k,l,x) P_{Y|X}(y|x) \sum_{\tilde{l}} \mathbf{1}[(u_{k\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \quad (317)$$

$$= \frac{1}{|\mathcal{M}|} \mathbb{E}_{\mathcal{C}} \left[\sum_{k,y} \frac{1}{|\mathcal{K}|} P_Y(y) \sum_{\tilde{l}} \mathbf{1}[(u_{k\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \quad (318)$$

$$= \frac{|\mathcal{L}|}{|\mathcal{M}|} \sum_{u,y} P_Y(y) P_U(u) \mathbf{1}[(u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \quad (319)$$

$$= \frac{|\mathcal{L}|}{|\mathcal{M}|} \sum_{(u,y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)} P_Y(y) P_U(u) \quad (320)$$

where (314) follows from the union bound and (319) follows from the fact that \mathcal{C} is generated according to P_U .

By uniting (311), (312), and (320) with (302), we have

$$\begin{aligned} & \mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbf{P}_e(\Phi; D)] \\ & \leq P_{UXY}((u, x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c) \cup (u, x, y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) \\ & \quad + \frac{|\mathcal{L}|}{|\mathcal{M}|} \sum_{(u,y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)} P_Y(y) P_U(u) + \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UX})}{|\mathcal{L}|}} + \frac{1}{2} \sqrt{\frac{2\gamma}{|\mathcal{K}||\mathcal{L}|}} + P_U[-\log P_U(u) > \gamma]. \end{aligned} \quad (321)$$

Since we can choose $\gamma > 0$ and $|\mathcal{K}|$ arbitrarily large, we have

$$\begin{aligned} & \mathbb{E}_\kappa \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbf{P}_e(\Phi; D)] \\ & \leq P_{UXY}((u, x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c) \cup (u, x, y) \notin \mathcal{T}_d^{\text{WZ}}(D) \cup (u, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) \\ & \quad + \frac{|\mathcal{L}|}{|\mathcal{M}|} \sum_{(u,y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)} P_Y(y) P_U(u) + \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{UX})}{|\mathcal{L}|}} + \delta. \end{aligned} \quad (322)$$

Consequently, there exists at least one (k, κ, \mathcal{C}) such that $\mathbf{P}_e(\Phi; D)$ is smaller than the right-hand-side of the inequality above. This completes the proof of Theorem 17.

APPENDIX G

PROOF OF THE NON-ASYMPTOTIC BOUND FOR GP IN THEOREM 19

A. Code Construction

As in WAK, we use the stochastic map introduced in Appendix C. In GP coding, let \mathcal{K} be the alphabet of the common randomness for the encoder and the decoder, and let $\mathcal{Z} = \mathcal{S}$ and $P_{UZ} = P_{US}$. Note that that, $\mathcal{T}_c(\gamma_c)$ defined in (203) is equivalent to $\mathcal{T}_c^{\text{GP}}(\gamma_c)$ defined in (66) in this case.

For GP coding, we construct $|\mathcal{M}|$ stochastic maps. Each stochastic map corresponds to a message in \mathcal{M} . For each message $m \in \mathcal{M}$, generate a codebook $\mathcal{C}^{(m)} = \{u_{11}^{(m)}, \dots, u_{|\mathcal{K}||\mathcal{L}|}^{(m)}\}$ where each $u_{kl}^{(m)}$ is independently drawn according to P_U . Then, for each $\mathcal{C}^{(m)}$ ($m \in \mathcal{M}$), construct a stochastic map $\varphi_{\mathcal{C}^{(m)}}$ as defined in (230).

By using $\{\varphi_{\mathcal{C}^{(m)}}\}_{m \in \mathcal{M}}$, we construct a GP code Φ as follows. Given the message $m \in \mathcal{M}$, the channel state $s \in \mathcal{S}$, and the common randomness $k \in \mathcal{K}$, the encoder first generates $l \in \mathcal{L}$ according to $\varphi_{\mathcal{C}^{(m)}}(\cdot | k, s)$. Then, the encoder generates $x \in \mathcal{X}$ according to $P_{X|US}(\cdot | u_{kl}^{(m)}, s)$ and inputs x into the channel. If the randomly generated x results in $g(x) > \Gamma$ (i.e., the channel input does not satisfy the cost constraint), declare a cost-constraint violation error.⁷ Given the channel output $y \in \mathcal{Y}$ and the common randomness $k \in \mathcal{K}$, the decoder finds the unique index $\hat{m} \in \mathcal{M}$ such that

$$(u_{kl}^{(\hat{m})}, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p) \quad (323)$$

for some $l \in \mathcal{L}$. If there is no unique index $\hat{m} \in \mathcal{M}$ or more than one, declare a decoding error. This is a Feinstein-like decoder [7] for average probability of error. If no such unique \hat{m} exists, or if there exists more than one such \hat{m} , then a decoding error is declared.

⁷Even if $g(x) > \Gamma$ occurs, we still send x through the channel. The error event for this occurrence is accounted for in (326).

B. Analysis of Error Probability

Without loss of generality, and by symmetry, we may assume that $M = 1$ is the message sent. Let \hat{L} be the random index chosen by the helper via the stochastic map $\varphi_{\mathcal{C}^{(1)}}(\cdot | K, S)$, and let $\hat{U} = u_{K\hat{L}}$ be the chosen codeword. Note that the joint distribution of K, \hat{L}, \hat{U}, S is given as follows; cf. (231)

$$P_{K\hat{L}\hat{U}S}(k, l, u, s) = \frac{1}{|\mathcal{K}|} P_S(s) \varphi_{\mathcal{C}^{(1)}}(l | k, s) \mathbf{1}[u_{kl}^{(1)} = u] \quad (324)$$

and then, the joint distribution of $K, \hat{L}, \hat{U}, S, X, Y$ is given as

$$P_{K\hat{L}\hat{U}SXY}(k, l, u, s, x, y) = P_{K\hat{L}\hat{U}S}(k, l, u, s) P_{X|US}(x | u, s) W(y | x, s). \quad (325)$$

The smoothed versions $\bar{P}_{K\hat{L}\hat{U}S}$ and $\bar{P}_{K\hat{L}\hat{U}SXY}$ are given by substituting P_S in (324) with \bar{P}_S ; cf. (233).

If either a cost-constraint violation or a decoding error occurs, at least one of the following error events must occur:

$$\mathcal{E}_0 := \{x \notin \mathcal{T}_g^{\text{GP}}(\Gamma)\} \quad (326)$$

$$\mathcal{E}_1 := \{(u_{kl}^{(1)}, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)\} \quad (327)$$

$$\mathcal{E}_2 := \{\exists \tilde{m} \neq 1 \text{ s.t. } (u_{kl}^{(\tilde{m})}, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p) \text{ for some } l \in \mathcal{L}\}. \quad (328)$$

Hence, the error probability averaged over the all random codebook $\mathcal{C} := \{\mathcal{C}^{(m)}\}_{m \in \mathcal{M}}$ and the common randomness K can be bounded as

$$\mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbb{P}_e(\Phi; \Gamma)] \leq \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}SXY}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2)] \quad (329)$$

$$\leq \mathbb{E}_{\mathcal{C}^{(1)}} [P_{K\hat{L}\hat{U}SXY}(\mathcal{E}_0 \cup \mathcal{E}_1)] + \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}SXY}(\mathcal{E}_2)]. \quad (330)$$

At first, we evaluate the first term in (330). For a fixed codebook \mathcal{C} ,

$$\begin{aligned} & P_{K\hat{L}\hat{U}SXY}(\mathcal{E}_1) \\ &= \sum_{k, l, u, s, x, y} P_{K\hat{L}\hat{U}S}(k, l, u, s) P_{X|US}(x | u, s) W(y | x, s) \mathbf{1}[x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)] \end{aligned} \quad (331)$$

$$= \sum_{u, s, x, y} P_{\hat{U}S}(u, s) P_{X|US}(x | u, s) W(y | x, s) \mathbf{1}[x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)] \quad (332)$$

$$= P_{\hat{U}SXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)) \quad (333)$$

$$\leq \bar{P}_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)) + \frac{1 - \bar{P}_{USXY}(\mathcal{U} \times \mathcal{S} \times \mathcal{X} \times \mathcal{Y})}{2} + d(P_{\hat{U}SXY}, \bar{P}_{USXY}) \quad (334)$$

$$= \bar{P}_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)) + \frac{P_{USXY}((u, s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c))}{2} + d(P_{\hat{U}SXY}, \bar{P}_{USXY}) \quad (335)$$

$$\begin{aligned} &= \bar{P}_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)) + \frac{P_{USXY}((u, s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c))}{2} \\ &\quad + d(P_{\hat{U}S} P_{X|US} W, \bar{P}_{US} P_{X|US} W) \end{aligned} \quad (336)$$

$$\leq \bar{P}_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)) + \frac{P_{USXY}((u, s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c))}{2} + d(P_{\hat{U}S}, \bar{P}_{US}) \quad (337)$$

$$\leq \bar{P}_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p)) + P_{USXY}((u, s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c)) + d(\bar{P}_{\hat{U}S}, \bar{P}_{US}) \quad (338)$$

$$\begin{aligned} &= P_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p) \cap (u, s) \in \mathcal{T}_c^{\text{GP}}(\gamma_c)) + P_{USXY}((u, s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c)) \\ &\quad + d(\bar{P}_{\hat{U}S}, \bar{P}_{US}) \end{aligned} \quad (339)$$

$$= P_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u, y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p) \cup (u, s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c)) + d(\bar{P}_{\hat{U}S}, \bar{P}_{US}) \quad (340)$$

where (334) follows from (183), (337) follows from the data-processing inequality (182), and (338) follows from Lemma 35. By taking average over $\mathcal{C}^{(1)}$, the second term in (340) is upper bounded

$$\mathbb{E}_{\mathcal{C}^{(1)}} [d(\bar{P}_{\hat{U}S}, \bar{P}_{US})] \leq \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{US})}{|\mathcal{L}|}} + \frac{1}{2} \sqrt{\frac{2\gamma}{|\mathcal{K}||\mathcal{L}|}} + P_U[-\log P_U(u) > \gamma] \quad (341)$$

by using Lemma 36.

Next, we evaluate the second term in (330).

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} [P_{K\hat{L}\hat{U}SXY}(\mathcal{E}_2)] \\ &= \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,u,s,x,y} P_{K\hat{L}\hat{U}SXY}(k,l,u,s,x,y) \mathbf{1}[\exists \tilde{m} \neq 1, \tilde{l} \in \mathcal{L} \text{ s.t. } (u_{k\tilde{l}}^{(\tilde{m})}, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)] \right] \end{aligned} \quad (342)$$

$$\leq \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,u,s,x,y} P_{K\hat{L}\hat{U}SXY}(k,l,u,s,x,y) \sum_{\substack{\tilde{m} \neq 1 \\ \tilde{l} \in \mathcal{L}}} \mathbf{1}[(u_{k\tilde{l}}^{(\tilde{m})}, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)] \right] \quad (343)$$

$$\leq \mathbb{E}_{\mathcal{C}} \left[\sum_{k,l,u,s,x,y} P_{K\hat{L}\hat{U}SX}(k,l,u,s,x) W(y|x,s) \sum_{\tilde{m}, \tilde{l}} \mathbf{1}[(u_{k\tilde{l}}^{(\tilde{m})}, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)] \right] \quad (344)$$

$$= \mathbb{E}_{\mathcal{C}} \left[\sum_{k,y} \frac{1}{|\mathcal{K}|} P_Y(y) \sum_{\tilde{m}, \tilde{l}} \mathbf{1}[(u_{k\tilde{l}}^{(\tilde{m})}, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)] \right] \quad (345)$$

$$= |\mathcal{M}| |\mathcal{L}| \sum_{u,y} P_Y(y) P_U(u) \mathbf{1}[(u,y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)] \quad (346)$$

$$= |\mathcal{M}| |\mathcal{L}| \sum_{(u,y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)} P_Y(y) P_U(u) \quad (347)$$

where (346) follows from the fact that \mathcal{C} is generated according to P_U .

By combining (340), (341), and (347) with (330), we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbf{P}_e(\Phi; \Gamma)] \\ & \leq P_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u,y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p) \cup (u,s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c)) + |\mathcal{M}| |\mathcal{L}| \sum_{(u,y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)} P_Y(y) P_U(u) \\ & \quad + \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{US})}{|\mathcal{L}|}} + \frac{1}{2} \sqrt{\frac{2\gamma}{|\mathcal{K}| |\mathcal{L}|}} + P_U[-\log P_U(u) > \gamma] \end{aligned} \quad (348)$$

Since we can choose $\gamma > 0$ and $|\mathcal{K}|$ arbitrarily large, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}} \mathbb{E}_K [\mathbf{P}_e(\Phi; \Gamma)] \\ & \leq P_{USXY}(x \notin \mathcal{T}_g^{\text{GP}}(\Gamma) \cup (u,y) \notin \mathcal{T}_p^{\text{GP}}(\gamma_p) \cup (u,s) \notin \mathcal{T}_c^{\text{GP}}(\gamma_c)) + |\mathcal{M}| |\mathcal{L}| \sum_{(u,y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)} P_Y(y) P_U(u) \\ & \quad + \frac{1}{2} \sqrt{\frac{\Delta(\gamma_c, P_{US})}{|\mathcal{L}|}} + \delta. \end{aligned} \quad (349)$$

Consequently, there exists at least one realization of the random code (k, κ, \mathcal{C}) such that $\mathbf{P}_e(\Phi; \Gamma)$, defined in (25), is no larger than the right-hand-side of inequality (349). This completes the proof of Theorem 19.

APPENDIX H

PRELIMINARIES FOR PROOFS OF THE SECOND-ORDER CODING RATE

In this appendix, we provide some technical results that will be used in Appendices I, K, and M. More specifically, we will use the following multidimensional Berry-Esséen theorem and its corollary.

Theorem 37 (Göetze [23]). *Let $\mathbf{U}_1, \dots, \mathbf{U}_n$ be independent random vectors in \mathbb{R}^k with zero mean. Let $\mathbf{S}_n = \frac{1}{\sqrt{n}}(\mathbf{U}_1 + \dots + \mathbf{U}_n)$, $\text{Cov}(\mathbf{S}_n) = \mathbf{I}$, and $\xi = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\mathbf{U}_i\|_2^3]$. Let the standard Gaussian random vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, for all $n \in \mathbb{N}$, we have*

$$\sup_{\mathcal{C} \in \mathcal{C}_k} |\Pr\{\mathbf{S}_n \in \mathcal{C}\} - \Pr\{\mathbf{Z} \in \mathcal{C}\}| \leq \frac{254\sqrt{k}\xi}{\sqrt{n}}, \quad (350)$$

where \mathfrak{C}_k is the family of all convex, Borel measurable subsets of \mathbb{R}^k .

It should be noted that Theorem 37 can be applied for random vectors that are independent but not necessarily identical. For i.i.d. random vectors, Bentkus [52] proved that the dependency of the bound on the dimension can be improved from \sqrt{k} to $d^{1/4}$.

We will frequently encounter random vectors with non-identity covariance matrices. Thus, we slightly modify Theorem 37 in a similar manner as [25, Corollary 7] as follows.

Corollary 38. *Let $\mathbf{U}_1, \dots, \mathbf{U}_n$ be independent random vectors in \mathbb{R}^k with zero mean. Let $\mathbf{S}_n = \frac{1}{\sqrt{n}}(\mathbf{U}_1 + \dots + \mathbf{U}_n)$, $\text{Cov}(\mathbf{S}_n) = \mathbf{V} \succ 0$, and $\xi = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\mathbf{U}_i\|_2^3]$. Let the Gaussian random vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$. Then, for all $n \in \mathbb{N}$,*

$$\sup_{\mathcal{C} \in \mathfrak{C}_k} |\Pr\{\mathbf{S}_n \in \mathcal{C}\} - \Pr\{\mathbf{Z} \in \mathcal{C}\}| \leq \frac{254\sqrt{k}\xi}{\lambda_{\min}(\mathbf{V})^{3/2}\sqrt{n}}, \quad (351)$$

where \mathfrak{C}_k is the family of all convex, Borel measurable subsets of \mathbb{R}^k , and where $\lambda_{\min}(\mathbf{V})$ is the smallest eigenvalue of \mathbf{V} .

APPENDIX I

ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR WAK IN THEOREM 24

Proof: It suffices to show the inclusion $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon)$ for fixed $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$.

We first consider the case such that $\mathbf{V} = \mathbf{V}(P_{UTXY}) \succ 0$. First, note that $\mathbf{R} \in \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY})$ implies

$$\tilde{\mathbf{z}} := \sqrt{n} \left(\mathbf{R} - \mathbf{J} - \frac{2 \log n}{n} \mathbf{1}_2 \right) \in \mathcal{S}(\mathbf{V}, \varepsilon). \quad (352)$$

We fix a time-sharing sequence $t^n \in \mathcal{T}^n$ with type $P_{t^n} \in \mathcal{P}_n(\mathcal{T})$ such that

$$|P_{t^n}(t) - P_T(t)| \leq \frac{1}{n} \quad (353)$$

for every $t \in \mathcal{T}$ [45]. Then, we consider the test channel given by $P_{U^n|Y^n}(u^n|y^n) = P_{U|TY}^n(u^n|t^n, y^n)$, and we use Corollary 15 for $P_{U^n X^n Y^n} = P_{XY}^n P_{U^n|Y^n}$ by setting $\gamma_b = \log |\mathcal{M}_n| - \log n$, $\gamma_c = \log |\mathcal{L}_n| - \log n$, and $\delta = \frac{1}{n}$. Then, there exists a WAK code Φ_n such that

$$1 - P_e(\Phi_n) \geq \Pr \left\{ \sum_{i=1}^n \mathbf{j}(U_i, X_i, Y_i|t_i) \leq n\mathbf{R} - \log n \mathbf{1}_2 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}} \quad (354)$$

$$= \Pr \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{j}(U_i, X_i, Y_i|t_i) - \mathbf{J}) \leq \tilde{\mathbf{z}} + \frac{\log n}{\sqrt{n}} \mathbf{1}_2 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}}. \quad (355)$$

By using Corollary 38 to the first term of (355), we have

$$1 - P_e(\Phi_n) \geq \Pr \left\{ \mathbf{Z} \leq \tilde{\mathbf{z}} + \frac{\log n}{\sqrt{n}} \mathbf{1}_2 \right\} - O \left(\frac{1}{\sqrt{n}} \right) \quad (356)$$

$$= \Pr\{\mathbf{Z} \leq \tilde{\mathbf{z}}\} + O \left(\frac{\log n}{\sqrt{n}} \right) \quad (357)$$

$$\geq 1 - \varepsilon \quad (358)$$

for sufficiently large n , where (357) follows from the Taylor's approximation, and (358) follows from (352).

Next, we consider the case with \mathbf{V} is singular but not 0. In this case, we cannot apply Corollary 38 because $\lambda_{\min}(\mathbf{V}) = 0$. Since $\text{rank}(\mathbf{V}) = 1$, we can write $\mathbf{V} = \mathbf{v}\mathbf{v}^T$ by using the vector \mathbf{v} . Let $\mathbf{A}_i = \mathbf{j}(U_i, X_i, Y_i|t_i) - \mathbf{J}$. Then we can write $\mathbf{A}_i = \mathbf{v}B_i$ by using the scalar independent random variables $\{B_i\}_{i=1}^n$. Thus, by using the ordinary Berry-Esséen theorem [63, Ch. XVI] for $\{B_i\}_{i=1}^n$, we can derive (358).

Finally, we consider the case where $\mathbf{V} = \mathbf{0}$. In this case, by setting $\tilde{\mathbf{z}} = \mathbf{0}$ in (355), we can find that the right hand side converges to 1. \blacksquare

APPENDIX J

ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR WAK IN THEOREM 25

Proof: We only provide a sketch of the proof because most of the steps are the same as Appendix I. The only modification is that we use Theorem 16 instead of Corollary 15 by setting $\gamma_b = \log |\mathcal{M}_n| - \rho\sqrt{n} - \log n$, $\gamma_c = \log |\mathcal{L}_n| + \rho\sqrt{n} - \log n$, $J_n = |\mathcal{L}_n|2^{\rho\sqrt{n}}$, and $\delta = \frac{1}{n}$. ■

APPENDIX K

ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR WZ IN THEOREM 27

Proof: It suffices to show the inclusion $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, g) \subset \mathcal{R}_{\text{WZ}}(n, \varepsilon)$ for fixed $(P_{UTXY}, g) \in \tilde{\mathcal{P}}(P_{XY})$. We assume that $\mathbf{V} = \mathbf{V}(P_{UTXY}, g) \succ 0$, since the case where \mathbf{V} is singular can be handled in a similar manner as Appendix I (see also [25, Proof of Theorem 5]).

First, note that $[R, D]^T \in \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, g)$ implies

$$\tilde{\mathbf{z}} := \sqrt{n} \left(\begin{bmatrix} -\frac{1}{n} \log \frac{L_n}{|\mathcal{M}_n|} \\ \frac{1}{n} \log L_n \\ D \end{bmatrix} - \mathbf{J} - \frac{2 \log n}{n} \mathbf{1}_3 \right) \in \mathcal{S}(\mathbf{V}, \varepsilon) \quad (359)$$

for some positive integer L_n . We fix a sequence $t^n \in \mathcal{T}^n$ satisfying (353) for every $t \in \mathcal{T}$. Then, we consider the test channel given by $P_{U^n|X^n}(u^n|x^n) = P_{U|TX}^n(u^n|t^n, x^n)$, and we use Corollary 18 for $P_{U^nX^nY^n} = P_{XY}^n P_{U^n|X^n}$ by setting $\gamma_p = \log \frac{L_n}{|\mathcal{M}_n|} + \log n$, $\gamma_c = \log L_n - \log n$, and $\delta = \frac{1}{n}$. Then, there exists a WZ code such that

$$1 - \text{P}_e(\Phi_n; D) \geq \Pr \left\{ \sum_{i=1}^n \mathbf{j}(U_i, X_i, Y_i|t_i) \leq \begin{bmatrix} -\log \frac{L_n}{|\mathcal{M}_n|} \\ \log L_n \\ nD \end{bmatrix} - \log n \mathbf{1}_3 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}} \quad (360)$$

$$= \Pr \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{j}(U_i, X_i, Y_i|t_i) - \mathbf{J}) \leq \tilde{\mathbf{z}} + \frac{\log n}{\sqrt{n}} \mathbf{1}_3 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}}. \quad (361)$$

Now the rest of the proof proceeds by using the multidimensional Berry-Esséen theorem as in (356) to (358) for the WAK problem. ■

APPENDIX L

ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR LOSSY SOURCE CODING IN THEOREM 29

We slightly modify a special case of Corollary 18 as follows, which will be used in both Appendices L-A and L-B.

Corollary 39. *For arbitrary distribution $Q_{\hat{X}} \in \mathcal{P}(\hat{\mathcal{X}})$, and for arbitrary constants $\gamma_c, \nu \geq 0$ and $\delta, \tilde{\delta} > 0$, there exists a lossy source code Φ with probability of excess distortion satisfying*

$$\text{P}_e(\Phi; D) \leq P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + \tilde{\delta} + \sqrt{\frac{2\gamma_c}{\tilde{\delta}|\mathcal{M}|}} + \delta + 2^{-\nu}. \quad (362)$$

Proof: As a special case of Corollary 18, we have

$$\text{P}_e(\Phi; D) \leq P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{P_{\hat{X}}(\hat{x})} > \gamma_c \text{ or } d(x, \hat{x}) > D \right] + \tilde{\delta} + \sqrt{\frac{2\gamma_c}{\tilde{\delta}|\mathcal{M}|}} + \delta, \quad (363)$$

where we set $\gamma_p = 0$ and $L = \tilde{\delta}|\mathcal{M}|$. We can further upper bound the first term of (363) as

$$P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{P_{\hat{X}}(\hat{x})} > \gamma_c \text{ or } d(x, \hat{x}) > D \right] \quad (364)$$

$$= P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} + \log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \gamma_c \text{ or } d(x, \hat{x}) > D \right] \quad (365)$$

$$\leq P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } \log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \nu \text{ or } d(x, \hat{x}) > D \right] \quad (366)$$

$$\leq P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + P_{\hat{X}X} \left[\log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \nu \right] \quad (367)$$

$$= P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + P_{\hat{X}} \left[\log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \nu \right] \quad (368)$$

$$\leq P_{\hat{X}X} \left[\log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + 2^{-\nu}. \quad (369)$$

This completes the proof. \blacksquare

Remark 6. By showing Corollary 39 directly instead of via Corollary 18, we can eliminate the residual term $\tilde{\delta}$.

A. Proof Based on the Method of Types

To prove Theorem 29 by the method of types, we use the following lemma.

Lemma 40 (Rate-Redundancy [27]). *Suppose that $R(P_X, D)$ is differentiable w.r.t. D and twice differentiable w.r.t. P_X at some neighbourhood of (P_X, D) . Let ε be given probability and let ΔR be any quantity chosen such that*

$$P_X^n [R(P_{x^n}, D) - R(P_X, D) > \Delta R] = \varepsilon + g_n, \quad (370)$$

where $g_n = O\left(\frac{\log n}{\sqrt{n}}\right)$. Then, as n grows,

$$\Delta R = \sqrt{\frac{\text{Var}(j(X, D))}{n}} Q^{-1}(\varepsilon) + O\left(\frac{\log n}{n}\right). \quad (371)$$

Note that the quantity $j(x, D)$ has an alternative representation as the derivative of $Q \mapsto R(Q, D)$ with respect to $Q(x)$ evaluated at $P_X(x)$; cf. (152).

We also use the following lemma, which is a consequence of the argument right after [64, Theorem 1].

Lemma 41. *For a type $q \in \mathcal{P}_n(\mathcal{X})$, suppose that $\left|\frac{\partial R(q, D)}{\partial D}\right| < C$ for a constant $C > 0$ in some neighbourhood of q . Then, there exists a test channel $V \in \mathcal{V}_n(\mathcal{Y}; q)$ such that*

$$\sum_{x, \hat{x}} q(x) V(\hat{x}|x) d(x, \hat{x}) \leq D \quad (372)$$

and

$$I(q, V) \leq R(q, D) + \frac{\tau}{n}, \quad (373)$$

where τ is a constant depending on C , $|\mathcal{X}|$, $|\hat{\mathcal{X}}|$, and D_{\max} .

Using Lemmas 40 and 41, we prove Theorem 29.

Proof: We construct a test channel $P_{\hat{X}^n|X^n}$ as follows. For a fixed constant $\tilde{\tau} > 0$, we set

$$\Omega_n = \left\{ q \in \mathcal{P}_n(\mathcal{X}) : \|P_x - q\|^2 \leq \frac{\tilde{\tau} \log n}{n} \right\}. \quad (374)$$

Since we assumed that $R(P_X, D)$ is differentiable w.r.t. D at P_X , the derivative is bounded over any small enough neighbourhood of P_X . In particular, it is bounded by some constant C over Ω_n for sufficiently large n . For each $q \in \Omega_n$, we choose test channel $V_q \in \mathcal{V}_n(\mathcal{Y}; q)$ satisfying the statement of Lemma 41. Then, we define the test channel

$$P_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \begin{cases} \frac{1}{|\mathcal{T}_{V_{P_{x^n}}}(x^n)|} & \text{if } \hat{x}^n \in \mathcal{T}_{V_{P_{x^n}}}(x^n) \\ 0 & \text{else} \end{cases} \quad (375)$$

for x^n satisfying $P_{x^n} \in \Omega_n$, and otherwise we define $P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)$ arbitrarily as long as the channel only outputs \hat{x}^n satisfying $d_n(x^n, \hat{x}^n) \leq D$. Let $P_q \in \mathcal{P}_n(\hat{\mathcal{X}})$ be such that

$$P_q(\hat{x}) = \sum_x q(x) V_q(\hat{x}|x). \quad (376)$$

Then, let $\tilde{P}_q^n \in \mathcal{P}(\hat{\mathcal{X}}^n)$ be the uniform distribution on \mathcal{T}_{P_q} . Furthermore, let $Q_{\hat{X}^n} \in \mathcal{P}(\hat{\mathcal{X}}^n)$ be the distribution given by

$$Q_{\hat{X}^n}(\hat{x}^n) = \sum_{q \in \Omega_n} \frac{1}{|\Omega_n|} \tilde{P}_q^n(\hat{x}^n). \quad (377)$$

We now use Corollary 39 for $P_X = P_X^n$, $P_{\hat{X}|X} = P_{\hat{X}^n|X^n}$, and $Q_{\hat{X}} = Q_{\hat{X}^n}$. Then, by noting that

$$d_n(x^n, \hat{x}^n) = \sum_{x, \hat{x}} P_{x^n}(x) V_{P_{x^n}}(\hat{x}|x) d(x, \hat{x}) > D \quad (378)$$

never occurs for the test channel $P_{\hat{X}^n|X^n}$, we have

$$P_e(\Phi_n; D) \leq P_{\hat{X}^n X^n} \left[\log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \gamma_c - \nu \right] + \tilde{\delta} + \sqrt{\frac{2^{\gamma_c}}{\tilde{\delta} |\mathcal{M}_n|}} + \delta + 2^{-\nu} \quad (379)$$

$$= P_{\hat{X}^n X^n} \left[\frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n} \right] + \sqrt{\frac{n 2^{\tilde{\gamma} n}}{|\mathcal{M}_n|}} + \frac{3}{n}, \quad (380)$$

where we set $\gamma_c = \tilde{\gamma} n$, $\tilde{\delta} = \delta = \frac{1}{n}$, and $\nu = \log n$. Furthermore, by noting that

$$Q_{\hat{X}^n}(\hat{x}^n) \geq \frac{1}{|\Omega_n|} \tilde{P}_q^n(\hat{x}^n) \quad (381)$$

for any $q \in \Omega_n$, we have

$$P_{\hat{X}^n X^n} \left[\frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n} \right] \quad (382)$$

$$\leq P_{\hat{X}^n X^n} \left[\frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n}, P_{x^n} \in \Omega_n \right] + P_{X^n}[P_{x^n} \notin \Omega_n] \quad (383)$$

$$\leq P_{\hat{X}^n X^n} \left[\frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n}, P_{x^n} \in \Omega_n \right] + \frac{2\tilde{\tau}}{n^2} \quad (384)$$

$$\leq P_{\hat{X}^n X^n} \left[\frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{\tilde{P}_{P_{x^n}}^n(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n} - \frac{|\mathcal{X}| \log(n+1)}{n}, P_{x^n} \in \Omega_n \right] + \frac{2\tilde{\tau}}{n^2}, \quad (385)$$

where (384) follows from [27, Lemma 2] and (385) follows from (381) and the fact that $|\Omega_n| \leq |\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}$.

Furthermore, we also have

$$\log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{\tilde{P}_{P_{x^n}}^n(\hat{x}^n)} = \log \frac{|\mathcal{T}_{P_{P_{x^n}}}|}{|\mathcal{T}_{V_{P_{x^n}}}(x^n)|} \quad (386)$$

$$= nI(P_{x^n}, V_{P_{x^n}}) + O(\log n). \quad (387)$$

Thus, for $\mu_n = O\left(\frac{\log n}{n}\right)$, we have

$$P_e(\Phi_n; D) \leq P_{\hat{X}^n X^n} [I(P_{x^n}, V_{P_{x^n}}) > \tilde{\gamma} - \mu_n, P_{x^n} \in \Omega_n] + O\left(\frac{1}{n}\right) + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} \quad (388)$$

$$\leq P_{\hat{X}^n X^n} \left[R(P_{x^n}, D) > \tilde{\gamma} - \mu_n - \frac{\tau}{n}, P_{x^n} \in \Omega_n \right] + O\left(\frac{1}{n}\right) + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} \quad (389)$$

$$\leq P_{\hat{X}^n X^n} \left[R(P_{x^n}, D) > \tilde{\gamma} - \mu_n - \frac{\tau}{n} \right] + O\left(\frac{1}{n}\right) + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} \quad (390)$$

$$\leq P_{X^n} \left[R(P_{x^n}, D) > \tilde{\gamma} - \mu_n - \frac{\tau}{n} \right] + O\left(\frac{1}{n}\right) + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}}. \quad (391)$$

Thus, by setting $\tilde{\gamma} = R(P_X, D) + \Delta R$, $\frac{1}{n} \log |\mathcal{M}_n| = \tilde{\gamma} + \frac{2 \log n}{n}$ and by using Lemma 40 (with $g_n = O\left(\frac{\log n}{\sqrt{n}}\right)$ being the residual terms in (391)), we have

$$R(n, \varepsilon; D) \leq R(P_X, D) + \sqrt{\frac{\text{Var}(j(X, D))}{n}} Q^{-1}(\varepsilon) + O\left(\frac{\log n}{n}\right) \quad (392)$$

for sufficiently large n , which implies the statement of the theorem. \blacksquare

B. Proof Based on the D -tilted Information

Let

$$\mathcal{B}_D(x^n) := \{\hat{x}^n : d_n(x^n, \hat{x}^n) \leq D\} \quad (393)$$

be the D -sphere, and let $P_{\hat{X}^*}$ be the output distribution of the optimal test channel of

$$\min_{P_{\hat{X}^*|X}} I(X; \hat{X}). \quad (394)$$

$\mathbb{E}[d(X, \hat{X})] \leq D$

To prove Theorem 29 by the D -tilted information, we use the following lemma.

Lemma 42 (Lemma 2 of [28]). *Under some regularity conditions, which are explicitly given in [28, Lemma 2] and satisfied by discrete memoryless sources, there exists constants $n_0, c, K > 0$ such that*

$$P_X^n \left[\log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} \leq \sum_{i=1}^n j(x_i, D) + C \log n + c \right] \geq 1 - \frac{K}{\sqrt{n}} \quad (395)$$

for all $n \geq n_0$, where $C > 0$ is a constant given by [28, Equation (86)].

Proof: We construct test channel $P_{\hat{X}^n|X^n}$ as

$$P_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \begin{cases} \frac{P_{\hat{X}^*}^n(\hat{x}^n)}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} & \text{if } \hat{x}^n \in \mathcal{B}_D(x^n) \\ 0 & \text{else} \end{cases}. \quad (396)$$

We now use Corollary 39 for $P_X = P_X^n$, $P_{\hat{X}|X} = P_{\hat{X}^n|X^n}$, $Q_{\hat{X}} = P_{\hat{X}^*}^n$, $\gamma_c = \tilde{\gamma}n$, $\tilde{\delta} = \delta = \frac{1}{n}$ and $\nu = \log n$. Then,

by noting that $d_n(x^n, \hat{x}^n) > D$ never occur for the test channel $P_{\hat{X}^n|X^n}$, we have

$$P_e(\Phi_n; D) \leq P_{\hat{X}^n|X^n} \left[\log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{P_{\hat{X}^*}^n(\hat{x}^n)} > \tilde{\gamma}n - \log n \right] + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \quad (397)$$

$$\leq P_{\hat{X}^n|X^n} \left[\log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} > \tilde{\gamma}n - \log n \right] + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \quad (398)$$

$$= P_X^n \left[\log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} > \tilde{\gamma}n - \log n \right] + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \quad (399)$$

$$\leq P_X^n \left[\sum_{i=1}^n j(x_i, D) > \tilde{\gamma}n - (C+1)\log n - c \right] \quad (400)$$

$$+ P_X^n \left[\log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} > \sum_{i=1}^n j(x_i, D) + C\log n + c \right] + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \quad (401)$$

$$\leq P_X^n \left[\sum_{i=1}^n j(x_i, D) > \tilde{\gamma}n - (C+1)\log n - c \right] + \frac{K}{\sqrt{n}} + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n}, \quad (402)$$

where (402) follows from Lemma 42. Thus, by setting $\tilde{\gamma} = \frac{1}{n} \log |\mathcal{M}_n| - \frac{2\log n}{n}$ and by applying the Berry-Esséen theorem [63], we have (392) for sufficiently large n , which implies the statement of the theorem. ■

APPENDIX M

ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR GP IN THEOREM 30

Proof: It suffices to show the inclusion $\mathcal{R}_{\text{in}}(n, \varepsilon, P_{TUSXY}) \subset \mathcal{R}_{\text{GP}}(n, \varepsilon)$ for a fixed $P_{TUSXY} \in \tilde{\mathcal{P}}(W, P_S)$. We assume that $\mathbf{V} = \mathbf{V}(P_{TUSXY}, g) \succ 0$, since the case where \mathbf{V} is singular can be handled in a similar manner as Appendix I (see also [25, Proof of Theorem 5]).

First, note that $[R, \Gamma]^T \in \mathcal{R}_{\text{in}}(n, \varepsilon, P_{TUSXY})$ implies that

$$\tilde{\mathbf{z}} := \sqrt{n} \left(\begin{bmatrix} \frac{1}{n} \log |\mathcal{M}_n| L_n \\ -\frac{1}{n} \log L_n \\ -\Gamma \end{bmatrix} - \mathbf{J} + \frac{2 \log n}{n} \mathbf{1}_3 \right) \in -\mathcal{S}(\mathbf{V}, \varepsilon) \quad (403)$$

for some positive integer L_n . We fix a sequence $t^n \in \mathcal{T}^n$ satisfying (353) for every $t \in \mathcal{T}$. Then, we consider the test channel and the input distribution given by $P_{U^n X^n | S^n}(u^n, x^n | s^n) = P_{UX|TS}^n(u^n, x^n | t^n, s^n)$, and we use Corollary 20 for $P_{U^n X^n S^n Y^n} = P_S^n P_{U^n X^n | S^n} W^n$ by setting $\gamma_p = \log |\mathcal{M}_n| L_n + \log n$, $\gamma_c = \log L_n - \log n$, and $\delta = \frac{1}{n}$. Then, there exists a GP code Φ_n whose average probability of error $P_e(\Phi_n; \Gamma)$ satisfies

$$1 - P_e(\Phi_n; \Gamma) \geq \Pr \left\{ \sum_{i=1}^n \mathbf{j}(U_i, S_i, X_i, Y_i | t_i) \geq \begin{bmatrix} \log |\mathcal{M}_n| L_n \\ -\log L_n \\ -n\Gamma \end{bmatrix} + \log n \mathbf{1}_3 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}} \quad (404)$$

$$= \Pr \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{j}(U_i, S_i, X_i, Y_i | t_i) - \mathbf{J}) \geq \tilde{\mathbf{z}} - \frac{\log n}{\sqrt{n}} \mathbf{1}_3 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}}. \quad (405)$$

Now the rest of the proof proceeds by using the multidimensional Berry-Esséen theorem as in (356) to (358) for the WAK problem. ■

REFERENCES

- [1] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge University Press, 2012.
- [2] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. on Inf. Th.*, vol. 21, no. 3, pp. 294–300, 1975.
- [3] R. Ahlswede and J. Körner, "Source coding with side information and a converse for the degraded broadcast channel," *IEEE Trans. on Inf. Th.*, vol. 21, no. 6, pp. 629–637, 1975.
- [4] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. on Inf. Th.*, vol. 22, no. 1, pp. 1–10, Jan 1976.

- [5] S. Gelfand and M. Pinsker, "Coding for channel with random parameters," *Prob. of Control and Inf. Th.*, vol. 9, no. 1, pp. 19–31, 1980.
- [6] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *Allerton Conference*, 2012.
- [7] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer Berlin Heidelberg, Feb 2003.
- [8] S. Miyake and F. Kanaya, "Coding theorems on correlated general sources," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer*, vol. E78-A, no. 9, pp. 1063–70, 1995.
- [9] K.-I. Iwata and J. Muramatsu, "An information-spectrum approach to rate-distortion function with side information," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer*, vol. E85-A, no. 6, pp. 1387–95, 2002.
- [10] V. Y. F. Tan, "The capacity of the general Gel'fand-Pinsker channel and achievable second-order coding rates," *arXiv:1210.1091*, Oct 2012.
- [11] M. Hayashi, "Second-order asymptotics in fixed-length source coding and intrinsic randomness," *IEEE Trans. on Inf. Th.*, vol. 54, pp. 4619–37, Oct 2008.
- [12] —, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. on Inf. Th.*, vol. 55, pp. 4947–66, Nov 2009.
- [13] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. on Inf. Th.*, vol. 39, no. 3, pp. 752–72, Mar 1993.
- [14] M. Hayashi, "General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel," *IEEE Trans. on Inf. Th.*, vol. 52, no. 4, pp. 1562–75, Apr 2006.
- [15] Z. Luo and I. Devetak, "Channel simulation with quantum side information," *IEEE Trans. on Inf. Th.*, vol. 55, no. 3, pp. 1331–1342, 2009.
- [16] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, "Entanglement-assisted classical capacity of noisy quantum channels," *Phys. Rev. Lett.*, vol. 83, no. 15, pp. 3081–3084, Oct 1999.
- [17] —, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Trans. on Inf. Th.*, vol. 48, no. 10, pp. 2637–2655, Oct 2002.
- [18] A. Winter, "Compression of sources of probability distributions and density operators," *arXiv:quant-ph/0208131*, 2002.
- [19] P. Cuff, "Distributed channel synthesis," *arXiv:1208.4415*, 2012.
- [20] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inf. Th.*, vol. 21, pp. 226–228, Mar. 1975.
- [21] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Inf. Th.*, vol. 19, pp. 471–80, 1973.
- [22] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. on Inf. Th.*, vol. 40, no. 4, pp. 1147–57, Apr 1994.
- [23] F. Göetze, "On the rate of convergence in the multivariate CLT," *The Annals of Probability*, vol. 19, no. 2, pp. 721–739, 1991.
- [24] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding in the finite blocklength regime," *IEEE Trans. on Inf. Th.*, vol. 56, pp. 2307–59, May 2010.
- [25] V. Y. F. Tan and O. Kosut, "On the dispersions of three network information theory problems," *arXiv:1201.3901*, Feb 2012, [Online].
- [26] D. Wang, A. Ingber, and Y. Kochman, "The dispersion of joint source-channel coding," in *Allerton Conference*, 2011, arXiv:1109.6310.
- [27] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *Data Compression Conference (DCC)*, 2011.
- [28] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. on Inf. Th.*, vol. 58, no. 6, pp. 3309–38, Jun 2012.
- [29] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Int. Conv. Rec.*, vol. 7, pp. 142?–163, 1959.
- [30] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- [31] R. Ahlswede and P. Gács and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, no. 3, pp. 157–177, 1976.
- [32] S. Kuzuoka, "A simple technique for bounding the redundancy of source coding with side information," in *Int. Symp. Inf. Th.*, Boston, MA, 2012.
- [33] B. Kelly and A. Wagner, "Reliability in source coding with side information," *IEEE Trans. on Inf. Th.*, vol. 58, no. 8, pp. 5086–5111, Aug 2012.
- [34] M. Costa, "Writing on dirty paper," *IEEE Trans. on Inf. Th.*, vol. 29, no. 3, pp. 439–441, Mar 1983.
- [35] P. Moulin and Y. Wang, "Capacity and random-coding exponents for channel coding with side information," *IEEE Trans. on Inf. Th.*, vol. 53, no. 4, pp. 1326–47, Apr 2007.
- [36] H. Tyagi and P. Narayan, "The Gelfand-Pinsker channel: Strong converse and upper bound for the reliability function," in *Proc. of IEEE Intl. Symp. on Info. Theory*, Seoul, Korea, 2009.
- [37] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. on Inf. Th.*, vol. 42, no. 1, pp. 63–86, Jan 1996.
- [38] P. Hayden, R. Jozsa, and A. Winter, "Trading Quantum for Classical Resources in Quantum Data Compression," *J. Math. Phys.*, vol. 43, no. 9, pp. 4404–4444, Sep 2002.
- [39] N. Datta, M.-H. Hsieh, and M. M. Wilde, "Quantum rate distortion, reverse shannon theorems, and source-channel separation," *IEEE Trans. on Inf. Th.*, vol. 59, no. 1, pp. 615–630, Jan. 2013.
- [40] M. H. Yassaee, M. R. Aref, and A. Gohari, "Achievability proof via output statistics of random binning," *arXiv:1203.0730*, 2012.
- [41] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie," in *Trans. Third. Prague Conf. Inf. Th.*, 1962, pp. 689–723.
- [42] I. Kontoyiannis, "Second-order noiseless source coding theorems," *IEEE Trans. on Inf. Th.*, pp. 1339–41, Jul 1997.
- [43] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, "How quickly can we approach channel capacity?" in *Asilomar Conf.*, 2004.
- [44] R. Nomura and T. S. Han, "Second-order resolvability, intrinsic randomness, and fixed-length source coding for mixed sources: Information spectrum approach," *IEEE Trans. on Inf. Th.*, vol. 59, no. 1, pp. 1–16, Jan 2013.
- [45] Y.-W. Huang and P. Moulin, "Finite blocklength coding for multiple access channels," in *Int. Symp. Inf. Th.*, 2012.
- [46] E. MolavianJazi and J. N. Laneman, "Simpler achievable rate regions for multiaccess with finite blocklength," in *Int. Symp. Inf. Th.*, Boston, MA, 2012.
- [47] R. Nomura and T. S. Han, "Second-order Slepian-Wolf source coding for mixed sources," *arXiv:1207.2505*, Jul 2012.

- [48] E. Haim, Y. Kochman, and U. Erez, "A note on the dispersion of network problems," in *Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, 2012.
- [49] A. Gupta and S. Verdú, "Operational duality between Gelfand-Pinsker and Wyner-Ziv coding," in *Intl. Symp. Inf. Th.*, Austin, TX, 2010.
- [50] A. Feinstein, "A new basic theorem of information theory," *IEEE Trans. on Inf. Th.*, vol. 4, no. 4, pp. 2–22, 1954.
- [51] E. A. Haroutunian, M. E. Haroutunian, and A. N. Harutyunyan, *Reliability Criteria in Information Theory and Statistical Hypothesis Testing*, ser. Foundations and Trends in Communications and Information Theory. Now Publishers Inc, 2008, vol. 4.
- [52] V. Bentkus, "On the dependence of the Berry-Esseen bound on dimension," *J. Stat. Planning and Inference*, vol. 113, pp. 385 ?– 402, 2003.
- [53] W. Gu, R. Koetter, M. Effros, and T. Ho, "On source coding with coded side information for a binary source with binary side information," in *Intl. Symp. Info. Th.*, Nice, France, July 2007.
- [54] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. on Inf. Th.*, vol. 29, no. 5, pp. 731–739, May 1983.
- [55] A. Ingber and M. Feder, "Finite blocklength coding for channels with side information at the receiver," in *Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, 2010.
- [56] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. on Inf. Th.*, vol. 31, no. 6, pp. 727–734, 1985.
- [57] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. on Inf. Th.*, vol. 28, no. 6, pp. 851–857, 1982.
- [58] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. on Inf. Th.*, vol. 25, pp. 306–311, Mar 1979.
- [59] A. El Gamal and E. C. van der Meulen, "A proof of Marton's coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. on Inf. Th.*, vol. 27, no. 1, pp. 120–122, Jan 1981.
- [60] T. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. on Inf. Th.*, vol. 25, no. 5, pp. 572–84, 1979.
- [61] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. on Inf. Th.*, vol. 33, no. 6, pp. 759–772, Jun 1987.
- [62] R. Renner and S. Wolf, "Simple and tight bounds for information reconciliation and privacy amplification," in *Advances in Cryptology—ASIACRYPT 2005, Lecture Notes in Computer Science, Springer-Verlag*, vol. 3788, Dec 2005, pp. 199–216.
- [63] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. John Wiley and Sons, 1971.
- [64] B. Yu and T. P. Speed, "A rate of convergence result for a universal d-semifaithful code," *IEEE Trans. on Inf. Th.*, vol. 39, no. 3, pp. 813–820, Mar 1997.